Cartographie des erreurs en anglais L2 : vers une typologie intégrant système et texte

Clive E. Hamilton

UNIVERSITE SORBONNE NOUVELLE - PARIS 3

École Doctorale 268 – Langage et Langues : description, théorisation, transmission

UMR 8094 – Langues, Textes, Traitements informatiques, Cognition (LATTICE)

Thèse de doctorat en sciences du langage

(VOLUME 1)

présentée par

Clive E. HAMILTON

CARTOGRAPHIE DES ERREURS EN ANGLAIS L2 : *VERS UNE TYPOLOGIE INTÉGRANT SYSTÈME ET TEXTE*

dirigée par

Shirley CARTER-THOMAS

Soutenue le 4 décembre 2015

Jury :

M. David BANKS :
Professeur émérite, Université de Bretagne Occidentale
Mme. Shirley CARTER-THOMAS :
Professeur des universités, Télécom Ecole de Management ; Université Sorbonne Nouvelle
Mme. Natalie KÜBLER :
Professeur des universités, Université Paris Diderot
M. Frédéric LANDRAGIN :
Directeur de recherche, CNRS ; Université Sorbonne Nouvelle
M. John OSBORNE :
Professeur des universités, Université de Savoie (Chambéry)

Résumé

L'objectif principal de ce travail est d'explorer la frontière entre les erreurs grammaticales d'une part et les erreurs textuelles d'autre part, dans les productions écrites des étudiants francophones rédigeant en anglais langue étrangère (L2) à l'université. Pour ce faire, un corpus de textes d'apprenants en anglais L2 a été recueilli et annoté par le biais de plusieurs schémas d'annotation. Le premier schéma d'annotation est issu de l'UAM CorpusTool, un logiciel qui fournit une taxonomie d'erreurs intégrée. Les premières annotations ont été croisées avec d'autres annotations issues des métafonctions sémantiques que nous avons établies, en nous appuyant sur la linguistique systémique fonctionnelle.

En plus de fournir des statistiques en termes de fréquence d'occurrence des erreurs spécifiques chez les apprenants francophones, le croisement des schémas a permis d'identifier certaines valeurs proprement phraséologique, sémantique et textuelle qui semblent poser des problèmes particulièrement épineux. A ce titre, une classification de ce que nous avons appelé des erreurs d'acceptabilité textuelle a été établie, dans le but notamment d'avoir une vue globale sur les erreurs identifiables à ce niveau d'analyse. En bref, le présent travail retrace donc le cheminement de l'ensemble de notre thèse de ses débuts conceptuels jusqu'à la proposition d'un modèle explicatif permettant d'établir la description de toute occurrence erronée identifiée en langue étrangère – qu'elle soit notamment grammaticale (c'est-à-dire, imputable au système linguistique) ou textuelle (c'est-à-dire, imputable au texte).

Mots clés : corpus d'apprenants, erreurs d'apprenant, linguistique systémique fonctionnelle, grammaticalité, acceptabilité, anglais de spécialité

Abstract

The main objective of this study is to try and pinpoint the frontier between grammatical (or sentence-level) errors on the one hand and textual errors on the other in university student essays. Accordingly, a corpus of English L2 learner texts, written by French learners, was collected and annotated using several annotation schemes. The first annotation scheme used is based on a model from the UAM CorpusTool software package, which provided us with an integrated error taxonomy. The annotations obtained were then cross-analyzed using the semantic metafunctions identified in systemic functional linguistics.

In addition to providing statistics in terms of specific error frequency, our cross analysis has identified some areas that appear to pose particularly difficult problems, i.e. phraseology, and certain semantic and textual constructions. A classification of what we have called textual acceptability errors has thus been established. In short, the thesis begins with an examination of conceptual issues and ends with the proposal for an explanatory model that can describe erroneous occurrences identified in a foreign language – whether they are grammatical (i.e., linked to the language system itself) or textual (i.e. linked to the text) in nature.

Keywords: learner corpus, learner errors, grammaticality, acceptability, English for Specific Purposes, systemic functional linguistics

Remerciements

Qu'il me soit en premier lieu permis de remercier Shirley Carter-Thomas, ma directrice de thèse, d'avoir accepté de diriger ce travail. Que Frédéric Landragin trouve également l'expression de ma reconnaissance d'avoir été mon co-encadrant. L'intérêt qu'ils ont tous les deux témoigné pour mes recherches, leurs disponibilités sans faille, leurs rigueurs scientifiques et nos échanges fréquents m'ont permis de cultiver un esprit critique et de porter ce projet à terme.

En second lieu, qu'il me soit permis de remercier David Banks, Natalie Kübler et John Osborne d'avoir accepté de faire partie de mon jury de thèse et de m'avoir ainsi témoigné leur confiance.

Plus personnellement, je voudrais remercier toutes les personnes qui m'ont apporté une aide technique ou matérielle notamment :

- les étudiants en première année de licence LSO de l'université Paris Dauphine (promotion, 2011) : d'avoir accepté de participer à l'étude
- Lise Fontaine, Cassandra Bartlett, Danielle McShine et Shirley Carter-Thomas : d'avoir accepté d'être des annotateurs indépendants
- Karolina Krawczak : d'avoir accepté de vérifier certains de mes tests et mes calculs
- Samuel Arrindell, Charmaine Gabriel, Christophe Naejus et tout particulièrement Bernard Frouin : d'avoir accepté de faire des relectures et d'apporter des suggestions/conseils

Enfin, une pensée particulière à mes co-équipiers du laboratoire LaTTiCe, mes amis et ma famille pour leurs soutien moral.

Sommaire

Liste des sigles, abréviation et conventions

ARCTA	Aide à la Rédaction de Textes en Anglais (ARCTA)
AE ; EA	Analyse d'erreurs ; Error analysis
ANGLISH corpus	Corpus de données comparatives de l'anglais lu, répété et parlé en L1 & L2
B1, B2 ; C1, C2	Approximativement niveau intermédiaire et avancé, selon le CECRL
CECRL	Cadre européen commun de référence pour les langues
CEFLE Corpus	Corpus Écrit de Français Langue Étrangère
Dire Autrement	Corpus de textes de types variés produits par les apprenants de français L2
EFL	English as a Foreign Language
FLLOC Corpus	French Learner Language Oral Corpora
FRIDA	French Interlanguage Database
IBLC	Indianapolis Business Learner Corpus
ICLE	International Corpus of Learner English
KWIC	key-word-in-context
L1 (LM), L2, L3	Langue maternelle, langue étrangère, troisième langue étrangère
LANSAD	Langues pour Spécialistes d'Autres Disciplines
LLC	London-Lund Corpus
LP	Learner Profile
LSF	linguistique systémique fonctionnelle
métafonctions	métafonctions sémantiques (? à garder, ou pas ?)
MeLLANGE	Multilingual eLearning in Language Engineering
MR.I – MR.IV	(Méta)règles de bonne formation textuelle
NLP	Natural language processing
POS	part-of-speech
SCG	Scale and Category Grammar
Scientext	Ensemble de quatre sous-corpus : 2 en anglais L1 & L2 ; 2 en français
SEU	Survey of English Use
TAL	Traitement automatique des langues
TREACLE	Teaching Resource Extraction from an Annotated Corpus of Learner English
T-unit	Minimal terminable units
UAM	UAM CorpusTool : le logiciel utilisé pour effectuer les annotations
USE corpus	Uppsala Student English Corpus
Volet 1	Premier volet d'annotation : recueil des erreurs du système linguistique
Volet 2	Deuxième volet d'annotation : recueil des erreurs textuelles
WriCLE	Written Corpus of Learner English

« * »	Signale le caractère erroné d'une occurrence à l'étude
« […] »	Signale le caractère interrompu de la phrase
« [sic] »	Signale une erreur qui n'est pas prise en compte dans l'analyse
« $...$ »	Signale **une** possibilité de correction
« $##$ »	Signale qu'une seule réponse ne peut être clairement inférée
«(??)»	Signale qu'un doute persiste vis-à-vis d'un mot dans la version numérisée. Le mot en question est retranscrit, suivi de deux points d'interrogation : « mot(??) » (cf. Annexe : A6)
« (????)»	Signale l'illisibilité totale d'un mot dans la version numérisée. Le mot n'est donc pas retranscrit. (cf. Annexe : A6)

Liste des figures

Liste des tableaux

Introduction

La rédaction en langue étrangère n'est pas une activité simple, en ce qu'elle ne se résume pas à l'utilisation de la somme des connaissances linguistiques acquises dans une langue cible. Pourtant, la grammaticalité est souvent mise en avant comme le seul obstacle que les apprenants de langue étrangère doivent dépasser s'ils veulent « parfaire » en quelque sorte l'utilisation de toute langue qui viendra après leur langue maternelle. Force est de constater cependant que l'acceptabilité textuelle d'un énoncé est rarement remise en question alors qu'elle s'avère tout aussi décisive dans la compréhension et la « fluidité » d'un message – et ce, dans toute situation de communication. De plus, la maîtrise du premier ne présuppose pas la maîtrise du second, mais peu d'études examinent ces deux aspects conjointement. Notre travail se donne alors comme objectif d'examiner ce que nous appelons des erreurs à la fois en termes de grammaticalité mais également en termes d'acceptabilité textuelle. Nous porterons, de ce fait, une attention particulière au dernier aspect qui constitue, à notre sens, un domaine prometteur d'investigation scientifique tant pour les études en acquisition des langues étrangères que pour les études portant sur des corpus d'apprenants.

0.1 Objectif de notre de thèse

Partant d'un constat mis en avant dans un premier travail (Hamilton, 2011) démontrant que certains problèmes identifiés dans les écrits en langue étrangère ne résultent pas simplement d'une non-maîtrise linguistique, notre but est de faire émerger dans la présente étude une différenciation significative entre les erreurs dites grammaticales et ce que nous appelons les *erreurs d'acceptabilité textuelle*. Il est donc question d'adopter une perspective proprement textuelle où l'étude porte sur l'objet TEXTE, et par conséquent où les énoncés ne sont pas étudiés de manière « isolée » ou « décontextualisée ». Cela suppose, de ce fait, entrer dans une démarche à la fois descriptive et typologique, où il s'agirait d'identifier des problèmes, soit à un niveau local, soit à un niveau global, pour ensuite décrire leur spécificité qui provoque non seulement des erreurs imputables au système linguistique lui-même, mais également des erreurs qui génèrent des problèmes de compréhension du texte.

Précisons qu'une erreur locale est entendue dans le présent travail de façon lexico-grammaticale, à condition notamment qu'elle renvoie à la « grammaire de la phrase » par opposition à la « grammaire du texte » qui régit les erreurs que nous appelons textuelles (cf. section 1.3.3 et chapitre VI pour une définition élargie). A titre d'illustration, les deux premiers exemples ci-dessous correspondent aux erreurs locales et les deux derniers aux erreurs textuelles : en effet, les deux premiers exemples concernent le système linguistique, ou la grammaticalité de la phrase, tandis que, dans les deux autres exemples, la grammaticalité n'est pas affectée et nous n'avons visiblement pas affaire à des erreurs proprement locales. Les erreurs dans les exemples (iii) et (iv) ne peuvent donc être convenablement identifiées que si l'on prend connaissance des phrases autour de celles qui nous intéressent. Soulignons, de plus, pour ce qui est des erreurs textuelles, que le contexte plus large est souvent important pour se rendre compte de la spécificité de l'erreur. Mais dans un souci de brièveté nous avons fourni des exemples que l'on pourrait rapidement identifier sans trop de « co-texte » : au sens de pré et post-textes.

i. Nowadays, we can't go anywhere, watch anything on TV or listen to the radio without **see* $seeing$ or **heard* $hearing$ an ad or a commercial. (txt_046_sm1)
ii. However, […] in the society a gender gap which has significantly **rose* $risen$ **this* $these$ *past years*. (txt_060_sm1)
iii. In fact, to make money, a country has to produc[e] and consume or export. So **he* it needs to be competitive. Consequently **he* it cannot have a currency weaker than **his* its neighbours. **He* It has to be attractive and to be part of the international trade system. (txt_083_sm1)
iv. **Since its first day* $From day one$, the European Union has been growing step by step. Despite of a lot of difficulties *through years[sic]*, the EU still *exist[sic]* and *develop[sic]*itself now, thanks to so many political leaders. (txt_83_sm2)

Nos hypothèses initiales sur les erreurs de grammaticalité et de textualité en langue étrangère nous ont conduits à nous poser une série de questions - auxquelles il faudra répondre de façon individuelle avant de pouvoir établir une vue d'ensemble sur la problématique générale. Ces questions ont à leur tour conduit à la mise en œuvre des différentes variables - à la fois indépendante et dépendante (cf. chapitre IV) – à travers lesquelles nous avons l'intention de fournir une vue holistique sur l'ensemble des erreurs rencontrées dans notre corpus d'étude. Certaines de ces questions sont énumérées ci-dessous :

1. Si les connaissances grammaticales (à la fois en termes de compétence et de performance) sont la seule cause des erreurs, comment expliquer celles identifiées dans notre corpus d'apprenants intermédiaires[1] et avancés – notamment dans le chapitre VI ?
2. Où se situe la frontière entre les erreurs lexico-grammaticales et celle des erreurs dites textuelles ? Et est-ce que l'ajout d'un nouveau cadre théorique à l'analyse des erreurs (AE), telle que mise en œuvre par Corder (cf. section 1.2.2.2), permettra d'élargir le débat en fournissant de nouvelles perspectives à ce qui a été jugé dépourvu de cadre linguistique propre ?
3. Y a-t-il des différences significatives entres les erreurs observées, selon l'historique langagier d'un apprenant ?
4. Y a t-il une différence significative entre les diverses catégories d'erreurs textuelles?
5. Quel est l'impact d'un enseignement contextualisé, a posteriori non linguistique, sur les différentes catégories d'erreurs observées sur une année universitaire ?
6. Y a-t-il un risque de *fossilisation textuelle* si les erreurs ne sont pas élucidées, au temps opportun ?
7. Est-ce qu'une approche rigoureuse basée sur la linguistique de corpus permettra de mieux examiner nos deux types d'erreurs – à savoir, les erreurs du système linguistique (cf. chapitre V) et les erreurs textuelles (cf. chapitre VI), en fournissant des fréquences d'occurrence statistiquement significative pour tous les deux ?

Comme en témoigne la diversité de nos questions de recherche qui peuvent apparaître à première vue assez distinctes les unes des autres, il y a de nombreux enjeux à exploiter dans le présent travail de doctorat. Mais, au fur et à mesure que nous progressons dans les différents chapitres, nous allons démontrer que les questions de recherche ci-dessus sont toutes intrinsèquement liées à notre problématique et constitueront les « chaînons manquants » que l'on doit identifier si l'on veut pouvoir répondre, de manière avisée, aux questions portant sur les erreurs textuelles en langue étrangère.

Par ailleurs, comme nous l'avons souligné ci-dessus, la présente analyse se veut « contextualisée », en raison du fait que nous considérons que le contexte est à l'acceptabilité, ce que la grammaire est à la grammaticalité. Autrement dit, le contexte peut influencer l'acceptabilité d'un énoncé au même titre que la grammaire sur une phrase donnée. De la même façon, notre projet s'inscrit dans une approche à la fois sémantique et textuelle, puisant essentiellement dans des théories de la linguistique susceptibles de fournir un cadre d'analyse où le contexte est considéré comme partie intégrante de l'analyse. Ceci est le cas de la linguistique systémique fonctionnelle (LSF, cf. chapitre III) de Halliday & Matthiessen (2004) et, à ce titre, nous l'adoptons comme cadre d'analyse

[1]Ce niveau correspond approximativement à B1-B2 dans l'échelle du cadre européen commun de référence pour les langues (CECRL).

linguistique principale. Il en va sans dire que nous adopterons également des terminologies proprement LSF afin d'expliquer certains phénomènes observés tout au long de ce travail.

L'utilisation de la linguistique systémique fonctionnelle apporte plusieurs avantages à notre projet (cf. section 3.4 pour une discussion détaillée), notamment en nous permettant d'apporter un angle novateur à l'analyse traditionnelle des erreurs (AE). En effet l'AE a été souvent critiquée dans le sens qu'elle ne permet pas de rendre compte ou d'expliquer l'ensemble des erreurs rencontrées chez les apprenants (James, 1998). C'est-à-dire, le cadre général de l'analyse des erreurs ne différencie pas les erreurs selon qu'elles sont de type lexical, phrastique ou appartenant à une catégorie supérieure : toute l'analyse est basée sur la grammaire traditionnelle et tout doit rentrer par conséquent dans des classifications grammaticales « de surface ». La LSF nous permet donc d'élargir le spectre d'analyse en facilitant l'étude de toute unité erronée, qu'elle soit de type lexical individuel, multi-mots voire proprement phrastique. De surcroît, le cadre systémique permet également de s'intéresser aux erreurs dont le contexte est l'élément clé pour l'analyse. Par exemple dans le cas des analyses effectuées sur des empans textuels plus larges : à savoir pour des questions de progression thématique ou de manquements dits de cohérence ou de cohésion.

Outre le cadre de la linguistique systémique fonctionnelle, nous nous appuyons sur les travaux issus de deux lignées de recherche distinctes. D'une part, pour le repérage des erreurs, tous types confondus, notre analyse s'inscrit dans les débuts conceptuels des travaux de linguistes tels que Frei (1929) et Corder (1967), tout en prenant en compte des avancées méthodologiques développées d'autre part par O'Donnell (2008), O'Donnell et al. (2009) et Granger (2002, 2003, 2008). La notion d'erreur textuelle sera alors explicitée en prenant comme point de départ les travaux de ceux qui étudient les phénomènes que l'on pourrait qualifier de façon neutre comme « des dysfonctionnements textuels ». Une sélection de travaux qui ont guidé et façonné notre réflexion et qui s'inscrit de manière globale dans les lignées périphériques aux auteurs précédemment cités sera également présentée dans les deux premiers chapitres. Mais ajoutons ici à titre d'information que l'approche textuelle principale que nous adoptons s'inspire tout particulièrement des travaux portant sur le discours et les problèmes liés à la connexité ou la continuité de la distribution informationnelle dans le texte (Charolles, 1978, 1989) ; et plus singulièrement les différents problèmes rencontrés dans les productions écrites des apprenants d'anglais langue étrangère au niveau proprement textuel (Carter-Thomas 1999a, 1999b, 2000).

En définitive, le présent travail vise à une analyse approfondie des productions langagières en anglais langue étrangère – en s'intéressant à l'ensemble des écueils rencontrés par les apprenants

afin de mieux comprendre les limites que ceux-ci auront s'ils se basent uniquement sur leurs connaissances grammaticales pour créer un ensemble textuel cohérent. Cet ensemble, précisons-le, est entendu ici en tant que texte argumentatif convenablement construit répondant à la fois aux exigences du code linguistique et à un genre textuel donné.

Signalons ici, en guise de conclusion de cette première section, que l'analyse des erreurs (et par voie de conséquence, notre travail de thèse) se veut un outil méthodologique très utile à la fois aux chercheurs et aux enseignants en ce qu'elle fournit un regard holistique sur l'ensemble des phénomènes déjà maîtrisés et ceux qui restent à acquérir par les apprenants. Un corpus de textes annotés « pour des erreurs de types variés » constitue de ce fait une base de données extrêmement riche pour ceux qui souhaitent créer des outils spécifiques à l'attention des apprenants : par exemple des logiciels de correction grammaticale ou orthographique automatique (Kübler, 1995). Ce type de corpus annoté constitue également une ressource non-négligeable pour ceux qui construisent des dictionnaires de langue pour apprenants : en raison notamment du fait que le corpus permet de signaler de manière statistique et factuelle les éléments les plus problématiques qui nécessiteront une attention ou des précisions particulières. Nous espérons ainsi pouvoir permettre aux enseignants ou toute personne s'intéressant aux écueils de la production en anglais langue étrangère de mieux appréhender ces problèmes et particulièrement ceux qui sont « endémiques » à nos apprenants francophones. Enfin, cette étude nous permet d'explorer de près les frontières du processus d'acquisition d'une langue étrangère par le biais de « la performance effective » dans l'usage réel de celle-ci. L'ensemble de ces points seront explicités davantage dans la section ci-après.

0.2 Objectifs et enjeux de l'analyse des erreurs

L'analyse des erreurs se définit comme un outil méthodologique utile tant pour l'enseignant – et par conséquent les apprenants – que pour les linguistes qui s'intéressent à l'acquisition d'une langue, qu'elle soit maternelle ou étrangère. Elle peut être envisagée sous un angle synchronique, diachronique ou longitudinal – tout en ayant pour objectif principal d'identifier, décrire et expliquer les erreurs qui ont été commises. Synchronique si l'on veut connaitre les difficultés langagières que rencontre une population particulière à un moment donné. Par exemple, les erreurs commises en anglais langue étrangère par des étudiants en début de leur parcours universitaire. Ou a contrario, diachronique ou longitudinal si l'on peut suivre et comparer les erreurs commises à deux ou plusieurs périodes différentes (ce qui est notre cas de figure). Elle offre ainsi un cadre facilitant un véritable suivi de la progression des apprenants.

Ces études ont donc le mérite de fournir un regard global au linguiste, lui permettant de mesurer et de suivre de manière empirique les différents stades d'acquisition d'une langue donnée. Autrement dit, une étude longitudinale pourrait permettre par exemple d'observer l'ordre d'acquisition des différents phénomènes linguistiques – allant de l'acquisition simple du lexique individuel, à l'acquisition des phénomènes complexes de la syntaxe et des termes des règles combinatoires jusqu'à l'aptitude à créer un positionnement d'auteur propre en termes d'« *interpersonal stance* » (cf. Lancaster 2011 ; Wharton 2012).

D'un point de vue strictement didactique ces mêmes études apportent plusieurs contributions et éclaircissements. Nous détaillerons deux d'entre- elles ci-après. En classe de langue, elles constituent des données observables par l'enseignant permettant, de ce fait, de porter une attention particulière à ce qui a été acquis et ce qui ne l'a pas été. En effet, O'Donnell et al (2009) affirment que : « [to] organise the teaching of English as a Foreign Language (EFL), it is important to have a clear picture of the grammatical competence of the learners at each level of proficiency ». Cette observation peut conduire donc à la révision de certains éléments en cours ou à la réadaptation des séquences pédagogiques en fonction de la difficulté envisagée par un point précis. L'autre aspect didactique peut se traduire par ce que l'on appelle communément « analyse des besoins » (cf. West 1994).

En effet, cette deuxième lecture de l'analyse des erreurs permet aux concepteurs de manuels et d'examens scolaires de mieux cibler les difficultés des apprenants. Cela peut être fait de plusieurs manières : (i) dans la ligne traditionnelle de l'analyse des besoins, cette analyse pourrait faciliter la construction d'un programme institutionnel en fonction des besoins réels observés ; (ii) soit d'une part en multipliant les explications assorties d'exemples des points les plus épineux et d'autre part en réduisant le contenu de ce qui ne constitue pas un obstacle majeur en cours de langue ; (iii) soit simplement en dressant une liste des erreurs les plus communément commises afin d'attirer l'attention des apprenants sur « les pièges » à éviter (cf. par exemple Fitikides (2003) et Turton & Heaton (1996) pour des illustrations établissant des listes d'erreurs commises en anglais).

Certains travaillent également sur l'analyse des erreurs d'un point du vue du traitement automatique des langues (TAL) où l'idée tend vers l'automatisation tout d'abord de la reconnaissance des erreurs (cf. De Felice 2008 ; Leacock et al 2010) ou vers des corrections automatiques ou semi-automatiques (cf. Albert et al 2009 ; Anderson 2011 ; Kübler & Cornu 1994). Il convient toutefois de préciser qu'aucun logiciel n'a, à ce jour, fait preuve d'une fiabilité adéquate dans son traitement automatique, ce qui signifie que l'analyse des erreurs demeurera

encore quelque temps manuelle ou semi-automatique (avec des corrections obligatoires d'un annotateur humain pour les multiples imprécisions attestées dans ces logiciels). Mais, la poursuite d'AE sur des corpus annotés - lesquels pourraient être « intégrés » dans ces futurs logiciels - permettra de créer des bases de données suffisamment fiables pour accélérer le processus d'automatisation dans le repérage et la correction des erreurs.

0.3 Organisation de la thèse

Dans cette section, nous détaillerons les sept chapitres que comporte cette thèse. Tout d'abord, le **premier chapitre** intitulé « Genèse d'un courant épistémologique » se veut assez général. Il retrace de manière chronologique les développements des principales théories de l'analyse des erreurs (AE). L'accent est mis sur l'introduction de l'erreur comme un objet d'étude en sciences humaines, premièrement chez les spécialistes de la langue maternelle avant d'être à son tour étudiée en linguistique appliquée. Le chapitre se clôt en présentant (i) les multiples façons d'envisager les erreurs selon les différentes terminologies en vigueur et (ii) quelques unes des tendances actuelles dans l'AE.

Le **chapitre II** établit une liste des phénomènes étudiés en AE : y compris des études sur le lexique et la syntaxe mais tout particulièrement les unités qui dépassent les frontières de l'unité maximale de la grammaire traditionnelle, à savoir la phrase. L'introduction des paradigmes culturaux dans l'AE est également examinée en détail avant d'aborder des aspects proprement pragmatiques, informationnels et textuels. Une attention tout à fait particulière est accordée dans ce chapitre à l'influence de la culture sur les styles rédactionnels attendus en milieu universitaire et les éventuels effets ou problèmes que les apprenants d'une langue étrangère pourraient rencontrer en voulant transposer un genre textuel d'une langue-culture à une autre.

Le **chapitre III** présente notre cadre conceptuel d'analyse. Nous démontrons dans ce chapitre que si la langue constitue un « *system of meaning potential* » (Halliday & Matthiessen 2004), l'analyse des erreurs doit se préoccuper aussi bien des ressources linguistiques mises en œuvre pour réaliser l'unité phrastique que du potentiel sémantique visé : c'est-à-dire dépasser l'analyse lexico-grammaticale, stricto sensu. Cela étant, la théorie de la linguistique systémique fonctionnelle est décrite de manière à la différencier de la grammaire traditionnelle. Nous retraçons ensuite les trois niveaux sémantiques (ou plus précisément les métafonctions sémantiques) mis en avant par la grammaire systémique fonctionnelle qui nous permettent de penser les erreurs non pas seulement

selon leurs fonctions grammaticales mais également à travers le potentiel sémantique que les apprenants-scripteurs cherchent à construire.

S'ensuit alors le **chapitre IV** qui présente les grandes lignes du cadre méthodologique. Afin de rendre la lecture plus compréhensible ce chapitre a été divisé en trois parties.

1. La première aborde l'introduction de deux outils méthodologiques dans la tradition de l'AE : (i) il s'agit principalement d'abord de retracer les avancées en linguistique de corpus qui ont été intégrées dans la méthodologie du recueil des données exploitées en AE ; (ii) s'ensuit alors l'introduction de la linguistique outillée qui s'avère désormais indispensable pour traiter des données encore plus volumineuses, vis-à-vis des nouvelles tendances adoptées par ceux qui construisent des corpus d'apprenants. Nous retraçons donc les apports de ces deux outils à la tradition d'AE et à notre étude avant d'expliquer comment nous les avons incorporés dans notre travail.
2. La deuxième partie met en avant le corpus d'étude : de sa conception théorique, sa construction, et les différentes manipulations nécessaires pour le rendre exploitable au vu des outils présentés dans la première partie de ce chapitre. Nous précisons également dans cette partie les méthodes employées pour sélectionner les sujets-participants et le corpus de l'étude, sans oublier bien entendu le logiciel d'annotation « UAM CorpusTool » et les divers schémas d'annotations utilisés pour analyser le corpus final.
3. Enfin la troisième partie résume les différents tests d'accord inter-annotateurs qui ont été effectués dans notre étude dans le but de juger de la valider de l'ensemble de nos annotations.

Le **chapitre V** est le premier de deux chapitres portant sur les résultats et, à ce titre, se veut principalement descriptif. En effet, ce chapitre présente de manière synthétique l'ensemble des résultats obtenus dans le premier volet d'annotation, à savoir les erreurs relevées à partir des premiers schémas d'annotation. Ces erreurs sont dites du système linguistique dans la mesure où elles renvoient à des items dont la grammaticalité constitue principalement la mise en cause dans la sélection des occurrences. Les mêmes erreurs sont ensuite ré-annotées et réexpliquées selon des schémas issus des métafonctions systémiques (cf. section 4.2.3.2, pour plus de précision), permettant ainsi de fournir une perspective que nous pensons novatrice dans la manière de classifier et interpréter les erreurs. L'apport de ce changement de paradigme est également discuté (cf. aussi la section 8.4.2, pour un bilan circonstancié).

Le **chapitre VI** est le deuxième chapitre qui porte spécifiquement sur les résultats. Contrairement au chapitre précédent, il est principalement question ici des étiquetages obtenus dans les deux volets d'annotation dans lesquels on ne s'intéresse qu'aux items considérés comme des erreurs textuelles. C'est-à-dire les éléments qui ne peuvent être considérés comme étant des erreurs du système linguistique, mais uniquement en tant qu'items irrecevables par rapport au contexte textuel. Dans un premier temps, les erreurs textuelles issues du premier volet d'annotations sont passées en revue, ensuite les différents types d'erreurs d'acceptabilité textuelle observées dans le deuxième volet sont détaillés. Ces derniers sont alors explicités à la lumière des trois métafonctions systémiques.

Le **chapitre VII** permet de discuter des résultats présentés dans les deux chapitres précédents, en procédant notamment à une analyse croisée portant sur les trois profils linguistiques établis dans le document annexe A2 et les deux grands types d'erreurs observés dans les chapitres V et VI. Ce regard croisé facilite notamment l'étude de l'influence de trois des variables dépendantes essentielles à notre analyse : à savoir (i) l'influence des contacts linguistiques précoces ; (ii) l'influence des contacts linguistiques prolongés (c'est-à-dire, des séjours en pays anglophones) ; et enfin l'influence de la langue maternelle sur les erreurs signalées dans les deux volets d'annotation. Il sera ensuite question de faire des parallèles entres les résultats obtenus dans la présente étude et ceux obtenus dans des études antérieures. De plus, l'incidence de ces nombreux contacts linguistiques sera mise en relation avec les notions de transfert et d'interférence, sans oublier, bien entendu une discussion sur les aléas de l'emploi de ces deux termes dans la caractérisation des erreurs.

Le **chapitre VIII,** le dernier de notre étude, se divise en quatre parties. La première section présente de manière synthétique les principaux résultats de la présente étude. L'implication didactique de ces résultats est ensuite brièvement discutée. La deuxième section renvoie aux observations les plus notables de notre étude, en abordant notamment les erreurs de calculs sémantiques et les erreurs de mise en phrase. La troisième section passe en revue les principaux types d'erreurs examinés dans la présente étude, en s'intéressant notamment au flou terminologique qui existe entre l'emploi d'un même terme à travers quatre études différentes. A cet effet, nous présenterons un schéma permettant non seulement de hiérarchiser les erreurs identifiées dans notre corpus, mais également d'expliquer l'ensemble des erreurs rencontrées en langue étrangère – indépendamment du niveau (local ou global) d'occurrence. Enfin, la quatrième et dernière section s'intéresse (i) aux questions périphériques qui ont guidé notre réflexion tout au long de ce travail et (ii) à l'ensemble des points qui constituent des limites de la présente étude.

(Chapitre I) Genèse d'un courant épistémologique : focus sur l'unité lexicale

Ce chapitre repasse en revue l'élément principal sur lequel porte le présent travail de doctorat, à savoir les erreurs relevées dans les productions langagières chez les apprenants en langue étrangère. Il convient de ce fait de dresser le portrait de la notion d'erreur, en examinant sa conception à travers des études antérieures, afin de mieux la mettre en rapport avec le positionnement définitoire adopté dans notre travail. Il sera donc principalement question dans ce chapitre de retracer quelques travaux notables sur les erreurs, tant en sciences du langage que dans des sciences connexes de manière à mettre en exergue l'apport de ces études aux champs définis. A travers un examen minutieux de ces travaux, nous explorerons et expliquerons ensuite les différentes périodes témoignant d'une montée d'intérêt initial suivi du déclin relatif de ce qui est communément appelé aujourd'hui « l'analyse des erreurs ». A la suite de ceci, l'accent sera mis sur l'avènement de l'analyse des erreurs comme un objet de recherche linguistique à part entière.

Plus concrètement, il sera question d'examiner les études portant sur les erreurs en sciences du langage en s'intéressant singulièrement aux travaux qui ont eu un impact considérable à la fois sur la recherche en acquisition-apprentissage en langue maternelle (L1) ainsi qu'en langue étrangère (L2). Nous porterons également une attention particulière aux erreurs dites locales, à savoir les unités lexicales individuelles ou les unités multi-mots par opposition aux erreurs que nous appelons phrastiques ou textuelles (cf. sections 2.3.3 et 6.1). Les premières renvoient en principe aux erreurs de nature lexicale, orthographique ou grammaticale tandis que les dernières ont plutôt une incidence sur la macrostructure textuelle qui les compose – posant de ce fait une question d'acceptabilité.

1.1 L'étude de l'erreur en sciences humaines et sociales

Commençons l'examen de la notion d'erreur à travers l'expression latine suivante : *Errare humanum est, perseverare diabolicum*[2]. Si nous acceptons la première moitié de ce célèbre adage comme étant véridique – c'est-à-dire que les erreurs font partie inhérente de la nature humaine - nous devons également admettre que nous succombons tous à cette vérité philosophique à un moment ou un autre. Il s'ensuit donc que cela peut se matérialiser dans les différents aspects de nos vies, indépendamment bien entendu des différences individuelles, culturelles ou sociétales. La deuxième moitié de ce proverbe qui est souvent tronqué et moins connu signifie littéralement que persévérer dans l'erreur est l'œuvre du diable. Si l'on extrapole, cela signifie qu'une fois identifiée, il est illogique ou pervers de répéter la même erreur. En suivant ce raisonnement, l'erreur est donc naturelle et inévitable sous réserve qu'elle soit « la première ». Cependant, dès lors qu'elle est avérée ou sa source identifiée, il convient de faire en sorte de ne pas la reproduire. Cela étant dit, en poussant ce raisonnement plus loin, cette phrase qui est devenue aujourd'hui monnaie courante pourrait nous aider à mieux comprendre pourquoi l'étude des erreurs occupe une place tant privilégiée dans certains domaines de recherche où le dénominateur commun est celui des activités et des zones d'intervention humaines : à savoir en sciences humaines et sociales par excellence.

En effet, s'il est naturel ou « attendu » d'avoir à affaire à des erreurs dès que l'homme fait partie intégrante d'une étude, on peut en déduire que les sciences qui étudient les activités humaines rencontreront à un moment donné l'occurrence naturelle des erreurs – et ce, quelle que soient leurs typologies exactes. Qu'il soit question d'erreurs de production (dans le sens de l'*output* ou de la création au sens large) ou de manipulation (au sens d'intervention et d'interaction avec le monde réel), il est par conséquent tout à fait compréhensible que ces événements soient examinés – que ce soit dans un cadre de recherche strict où l'intérêt principal s'articulerait autour de l'identification et la description, ou dans un cadre plus empirique où l'intérêt résiderait vraisemblablement dans l'anticipation et par conséquent la prévention ou la correction des dites erreurs. Ce raisonnement pourrait également expliquer pourquoi les sciences sélectionnées ci-dessous voient toutes un intérêt particulier dans la poursuite scientifique de l'analyse des erreurs – que cela soit à petite ou grande échelle.

[2] Notons que ce proverbe et certaines de ses dérivatives ont été attribués à Sénèque, malgré quelques objections notables. Pour les dérivatives, en général la première moitié reste inchangée. Deux autres versions sont par exemple (i) […] in errare perseverare stultum « il est stupide de persévérer dans l'erreur » et (ii) […] ignoscere divinum « pardonner est divin ». Précisons à titre accessoire cependant que seule la première des trois relève d'une utilisation constante depuis les années 1800 (résultats obtenus par le N-gram de google), l'usage des deux autres est assez récent.

En sciences humaines et sociales :

- En psychologie, l'analyse des erreurs constitue un terrain multidimensionnel en ce qu'elle renvoie à des réalités différentes chez les uns et les autres. Notons cependant, selon le *Psychology Dictionary*,[3] qu'elle se définit comme « l'étude à la fois des facteurs humains et des facteurs de conception qui peuvent conduire à une erreur ». Cette définition a le mérite de souligner les deux dimensions que l'on retrouve dans les travaux de Reason (1991, 2000)[4] qui sont exploitées tant en psychologie qu'en médecine. A titre d'information, ce dernier rapporte que l'erreur est souvent conçue comme un manquement personnel – et donc moralement condamnable, ou due à des facteurs extérieurs généralement situationnels ayant contribué à la dite erreur.

- En linguistique appliquée, l'analyse des erreurs renvoie principalement à l'ensemble des méthodologies inspirées des travaux de S. P. Corder dans les années 1960. Elle vise l'identification, la description et l'explication des erreurs commises par des apprenants en langue étrangère. Ces erreurs sont dites de production, tant à l'oral qu'à l'écrit et leurs enjeux à la fois didactiques et linguistiques feront l'objet d'une présentation détaillée dans les sections 1.2.2 et 4.1.3. Pour des besoins de brièveté, précisons que l'analyse des erreurs se fait aussi bien en didactique des langues maternelles qu'en didactique des langues étrangères et a pour but l'examen approfondi du processus d'acquisition d'une langue donnée. Notons par ailleurs que ces analyses peuvent également porter sur toute période de la vie humaine, allant des enfants en bas âge aux adultes ayant déjà acquis certaines compétences linguistiques dans une ou plusieurs langue(s) donnée(s).

En sciences dites exactes :

- En médecine, Reinertsen (2000) et Weingart et al. (2000) abordent le sujet en termes d'épidémiologie des erreurs : soit en identifiant les pratiques à risque du praticien médical pour le premier, soit en identifiant les erreurs liées à l'administration de médicaments pour le second. Sans entrer dans le détail un peu trop éloigné de notre sujet de recherche, ces deux travaux appellent à plus de vigilance et préconisent une systématisation des analyses et un recueil des erreurs de manière à sensibiliser et ainsi réduire les nombreux risques encourus chez les praticiens.

[3] Selon les informations fournies sur le site www.psychologydictionary.org, ce dictionnaire est réalisé par des professionnels en psychiatrie et psychologie et se veut une référence fiable dans le domaine [dernière consultation le 07.08.2014]

[4] Notons renvoyons à ces travaux en raison du nombre de citations et renvois systématiques que nous avons pu constater dans la littérature de psychologie consultée.

- En mathématiques et en statistiques l'analyse des erreurs constitue un cadre d'analyse spécifique que nous n'avons pas la prétention de pouvoir expliquer finement ici. Pour une présentation détaillée, voir Dumont (1989) pour le premier et Taylor (1997) pour le second.

En méthodologie de recherche :

- D'un point de vue purement méthodologique, l'analyse des erreurs peut se faire sur les pratiques, les protocoles d'expérimentation ou encore le fonctionnement des outils ou équipements méthodologiques. Ici, le sens diffère des contextes précédemment cités, en renvoyant plutôt à la bonne calibration ou jaugeage des équipements ou outils d'expérimentation et d'analyse de façon à minimiser les erreurs d'interprétation. Autrement dit, une attention particulière est portée sur la conception et la manipulation des outils pour ne pas générer des erreurs humaines. Dans ce cas de figure, on distingue deux types d'erreurs dites systématiques ou aléatoires. Le premier renvoie à un problème de mesure, de fonctionnement ou d'échantillonnage qui est répété à travers l'ensemble d'une étude, introduisant de ce fait un biais fâcheux dans les résultats finaux et par conséquent dans l'interprétation de l'étude concernée. Le deuxième, c'est-à-dire les erreurs aléatoires, n'est relevé que ponctuellement et n'influe pas de manière statistiquement significative sur les résultats. Soulignons enfin qu'une bonne prise en compte de ces aspects permet de réduire ou d'augmenter la solidité ou la fiabilité d'une étude.

Bien que nous soyons conscients de la nature limitée de ces exemples, ils sont mis en avant ici pour illustrer le fait que l'étude des erreurs existe sous des formes et dimensions différentes selon les disciplines. Nous soulignons, à titre d'illustration, qu'elle est présente de la même manière en études biologiques, physiques et informatiques, pour ne citer que quelques-unes. Cela étant, elle demeure un objet récurrent relevé, tantôt dans un cadre proprement de recherche ou méthodologique, tantôt dans un cadre que l'on pourrait désigner comme conseil-prévention.

1.2 L'erreur comme objet linguistique

La linguistique est souvent succinctement définie comme l'étude scientifique du langage et en tant que telle ; elle ambitionne d'explorer tous les aspects du langage humain. On pourrait ajouter ici que son objectif sous-jacent est de procéder à des descriptions systématiques des différents phénomènes rencontrés. Ces « phénomènes » renvoient mais ne sont pas limités à l'étude générale des systèmes linguistiques (leur organisation structurelle et fonctionnelle internes), l'évolution du langage à travers les études diachroniques et synchroniques et la modélisation linguistique, pour

n'en nommer que quelques-uns. Toutefois l'étude du langage n'est pas restreinte aux théorisations abstraites mais est également préoccupée par la description des phénomènes linguistiques de la vie réelle, comme par exemple l'étude de l'utilisation réelle ou effective des langues. Cela étant, l'usage réel renvoie à l'ensemble des productions langagières d'un locuteur potentiel et celles-ci peuvent tout naturellement comporter des occurrences que l'on pourrait qualifier comme ne relevant pas du « bon usage » voire plus précisément de l'usage dit « normatif ». Ces occurrences - qu'elles en soient systématiques ou aléatoires – font partie de l'usage réel et effectif et constituent à ce titre un objet de recherche linguistique viable.

Toutefois, la notion d'erreur en linguistique est complexe car son acceptation varie chez les uns et les autres selon les courants et écoles de pensée auxquels ils s'identifient. A titre d'illustration, il peut être question d'une distinction entre la norme au sens normatif et l'usage réel qu'en font les locuteurs. Dans la grammaire normative une erreur est tout ce qui s'écarte d'une règle prescrite. Ceci peut donc relever des règles combinatoires de la langue en tant que système complexe ou du « bon usage » selon des grammairiens. A cet égard, on peut se mettre d'accord par exemple avec la position évoquée par Authier et Meunier, à propos d'une certaine normativité de la grammaire française.

> La grammaire française, traditionnellement, fait une large part à des jugements de type normatif portés sur certaines « fautes » ou « incorrections » par référence à l'idée que la langue est un système qui a ses exigences internes (tel énoncé « n'est pas conforme au génie de la langue ») aussi bien qu'à des règles du discours et à des valeurs intellectuelles ou morales qui masquent ou traduisent des valeurs de caractère implicitement social : « X est une faute, est incorrect, ne se dit pas, n'est pas français, n'est pas élégant, est vulgaire...; Y est inacceptable, n'est pas permis... etc. » (1972 : 49-50).

Bien que nous ne travaillions pas dans la présente étude sur la langue française, à proprement parler, cette citation a le mérite d'expliciter clairement ce à quoi correspond la normativité. En effet, cette conception à notre sens n'est pas propre à la grammaire française. Signalons cependant à ce sujet que les langues vivantes sont en constante évolution et ne sont de ce fait pas des ensembles statiques - les normes ne sont pas immuables et peuvent par conséquent être actualisées, en cas de besoin. Dans le cas, disons de l'introduction ou l'utilisation de nouveaux items lexicaux (issus de l'argot, des néologismes, des régionalismes, etc.) que ceux-ci peuvent tout d'abord être considérés comme des déviances ou des items impropres lors de leur première utilisation jusqu'à ce qu'ils deviennent des formes acceptées de par la fréquence d'emploi statistiquement significative dans le langage quotidien ou par leur introduction dans des textes ou documents officiels. De plus,

l'erreur peut dépasser le clivage de norme et usage pour entrer dans un débat plus clivant où celle-ci est envisagée en termes de grammaticalité et d'acceptabilité. De manière globale, le premier étant basé sur la correction grammaticale d'un énoncé et le second sur son adéquation vis-à-vis du contexte d'utilisation.

Dans les cas susmentionnés, l'identification d'une erreur se traduit nécessairement par une forme d'évaluation ou un jugement de valeur d'une utilisation spécifique du langage. Et à notre sens il est discutable de savoir qui peut évaluer ou juger de la correction ou l'adéquation d'un énoncé. De toute évidence, ce sujet épineux mérite des éclaircissements. Et ce, afin d'élucider davantage la conception de l'erreur pour l'ensemble du travail qui suit. Il est donc important de trouver une distinction claire, ou au moins fonctionnelle, entre ce qui a été considéré comme des erreurs dans la recherche linguistique antérieure et les nombreuses sous-divisions que le terme a subi. Nous allons donc nous intéresser à l'introduction et à l'évolution de la notion d'erreur dans les différents domaines de recherche linguistique, en commençant par porter une attention particulière aux premières études portant sur les erreurs en langue maternelle et ensuite celles portant sur les erreurs commises en langue étrangère, avant de donner une définition précise du terme. Nous définirons également à la fin de ce chapitre deux alternatives que l'on retrouve dans la littérature et qui sont souvent utilisées de manière interchangeable.

1.2.1 En langue maternelle

Comme nous l'avons souligné ci-dessus, la conception de l'erreur varie d'un linguiste à un autre, notamment pour tout ce qui dépasse les règles de combinaisons syntaxiques qui – signalons-le – dépendent étroitement du système linguistique lui-même. A ce titre, il sera possible de se mettre sans grande difficulté d'accord sur une erreur de type morphologique dans le cas d'une conjugaison d'un verbe en anglais, tandis que l'acceptabilité d'une forme d'un auxiliaire modal que l'on trouve de plus en plus aujourd'hui pourrait être source de discorde en termes de classification. Par exemple pour ce qui relève de la morphologie flexionnelle du verbe *'to lie'* (au sens d'être étalé ou se trouver), l'erreur du type *'*The papers lied all over his desk'* paraitrait évident pour un évaluateur ; tandis que *'*I should of $should have$ bought him flowers*' risque d'être moins facile à classer. Ces différences de conception ou ces variations peuvent être observées à travers la méthodologie conceptuelle employée dans ce qu'on appelle aujourd'hui « l'analyse des erreurs ». En effet, selon que l'on se place du côté de l'analyse normative ou au contraire de la correction grammaticale, le cadre d'analyse ou la taxonomie résultante nommant les erreurs risquent de ne pas converger – ce qui pose un problème tant pour la reproductibilité que la solidité de l'analyse.

De plus, il est également possible d'avoir des différences notables dans la façon dont les erreurs sont examinées en langue maternelle et en langue étrangère, ne serait-ce qu'en comparant les approches issues des premiers travaux répertoriés portant sur les erreurs. En effet dans la littérature que nous avons pu consulter, l'analyse des erreurs est apparue en langue maternelle avant de faire son apparition dans les études en langues secondes – et ce, dès le début des années 1900 (cf. Wilson 1909). Et un premier constat notable dans ces analyses réside dans le fait que celles-ci n'ont pas toutes été effectuées dans le même cadre ou ne sont pas toutes issues d'une recherche linguistique au sens moderne, mais plutôt d'une simple enquête pédagogique dans le cas de Wilson (1920).

Dans la sous-section qui suit, nous allons de ce fait nous concentrer sur la perspective fournie par les premières études d'analyse des erreurs où les locuteurs ont commis des erreurs dans leur langue maternelle. Nous reviendrons ensuite sur les contributions de ces premières études qui ont permis l'avancement de l'analyse des erreurs dans les différents contextes d'étude en langue maternelle et leurs éventuelles applications pédagogiques à l'époque.

1.2.1.1 Wilson (1909, 1920-29)

Le choix d'examiner les travaux de G. M. Wilson est un choix qui se veut ici avant tout pragmatique. En effet, dans un souci d'accessibilité ce choix résulte principalement du fait que des travaux antérieurs ne sont pas facilement disponibles ni en version papier ni en version électronique. Cela étant dit, une partie importante des travaux de Wilson a été numérisée facilitant de ce fait son exploitation encore aujourd'hui. Notons tout de même que ces travaux ne sont pas moins méritants mais au contraire font en réalité partie des études initiales ayant donné naissance à une série d'études sur les erreurs de langue - notamment en langue maternelle chez les écoliers. Il est donc important que nous gardions à l'esprit que les travaux de Wilson pourraient être considérés à la fois comme (i) représentatifs de quelque chose qui a commencé avant lui et qu'il a simplement poursuivi dans la même veine, (ii) ou plutôt que ses travaux étaient avant-gardistes et donc distincts de ce qui se faisait dans le domaine, à l'époque[5].

Cependant en dépit du fait que nous ne pouvons pas soutenir que telle ou telle personne a effectué la première étude de cas de ce que nous appelons « l'analyse des erreurs » en langue maternelle, nous pouvons néanmoins affirmer que les travaux de Wilson ont eu un impact notable à la fois sur le cadrage des études portant sur les erreurs et sur les applications pédagogiques qui en ont découlé. Pour mieux situer l'approche de Wilson en L1, nous allons donc présenter succinctement les trois

[5] Signalons que ces travaux renvoient aux années 1900-1920, dans un contexte institutionnel nord-américain.

principales publications de ce dernier qui ont contribué à l'émergence et la poursuite des études du même type, dans plusieurs établissements pendant une période de 20 ans.

1909 : la première publication

L'article de 1909[6] fait état d'une étude pilote menée à petite échelle, dans une école nord-américaine sous la direction de George M. Wilson. Vingt-quatre enseignants de l'établissement étaient chargés de recueillir et corriger les travaux oraux et écrits de leurs élèves, tout en répertoriant les erreurs observées pendant une période de deux semaines. Les élèves en question relevaient - dans le système éducatif américain - de *grade 1* à *grade 8*[7]. Notons toutefois que des informations supplémentaires sur les profils, ne serait-ce que démographique ou concernant le parcours linguistique, des élèves ou les conseils donnés aux enseignants par l'investigateur principal n'ont pas été fournis.

Les résultats obtenus ont été présentés sous forme d'une liste de mots les plus problématiques, en termes de fréquence d'occurrence et de classe grammaticale (à savoir verbe, adverbe ou pronom). Les autres erreurs ont été signalées comme *'miscellaneous'* (*comprendre*, divers). Il en ressort de cette étude la présence de 228 erreurs au total, mais ce qui a été jugé le plus significatif est le nombre de types différents. En effet, 69 types individuels différents ont été relevés et les 159 autres erreurs sont signalées en tant que répétition d'un des types préalablement identifiés. Par exemple, une erreur d'accord entre un sujet et un verbe est signalée comme étant un type précis, et chaque nouveau cas comme une répétition de ce même type.

Grade	VB	PR	AD	Di	Tt	*n*_type
1	108	9	0	3	120	12
2	18	9	0	0	27	10
3	40	1	5	2	48	24
4	1	0	0	0	1	1
5	7	1	2	0	10	5
6	3	8	1	0	12	8
7	2	1	0	6	9	8
8		1	0	0	1	1
Tt	179	30	8	11	228	69

Tableau 1 : Table d'erreurs de Wilson

Dans le tableau 1 « *VB* » signifie un problème de verbe, « *PR* » un problème de pronom, « *AD* » un adverbe, « *Di* » un problème divers, « *Tt* » le total et « *n_type* » renvoie aux types distincts

[6] L'ensemble des éléments présentés ici ont été rapportés soit dans Connelly (1935) soit par les archives de la revue North Carolina Education, septembre 1910 – faute d'accès à l'article dans son intégralité.

[7] A titre de comparaison, « grade 1 » correspond aux cours préparatoires (CP) à l'école primaire dans le système éducatif français, et « grade 8 » correspond à la quatrième (4ème) au collège.

d’erreurs signalées à un « grade » donné. Cela étant, en dépit de la simplicité apparente du tableau et du nombre réduit d’exemples, Wilson a réussi à démontrer que sur les 69 types d’erreurs signalées, 10 représentent à eux seuls 58% de l’ensemble des erreurs identifiées. Autrement dit 133 sur les 228 erreurs renvoient à la répétition de 10 types d’erreurs distinctes : statistiquement parlant cela signifie treize occurrences ou répétitions par type. Alors que les autres types étaient relativement bien plus aléatoires : à savoir 95 erreurs individuelles restantes pour 59 types différents et donc une récurrence de 1,6 fois par type précis. Malgré leur diminution en nombre, il a également été souligné que les erreurs des niveaux supérieurs comprenaient généralement les mêmes types que les années ou grades inférieur(e)s.

Au vu de ces résultats, Wilson en a conclu que la réduction des erreurs observées nécessiterait l’implémentation d’exercices systématiques quotidiens permettant dans un premier temps de réduire toutes les erreurs fréquentes dès les premiers stades de l’éducation primaire et par voie de conséquence d’endiguer également le phénomène de répétition observé dans les niveaux ou « grades » supérieurs. En fin de compte, l'approche innovante de l'étude, aussi bien que la clarté de ses résultats ont été reçues à l’époque de manière si favorable que l’expérience a été reconduite dans de nombreuses écoles dans des villes et états environnants.

1920 : La reproduction de l’étude dans plusieurs établissements

Le deuxième article de Wilson compare des études réalisées entre 1909 à 1920, dans lesquelles l'objectif portait sur l’identification des erreurs langagières produites par des enfants dans leur langue maternelle. Le document compare en effet cinq études ayant des paramètres conceptuels similaires à celle menée en 1909, de sorte que l'auteur a pu procéder à une comparaison des erreurs identifiées dans chacune des études avec celles obtenues dans la sienne. Cependant la méthodologie de l’article appelle plusieurs remarques. Tout d’abord, l’auteur a procédé à une comparaison entre les différents types d’erreurs et les chiffres observés sans dresser le portrait des différences méthodologiques ou contextuelles qui existaient entre les études (taille de la population, niveaux des élèves, etc.). Ce qui est souligné par conséquent dans son nouvel article se résume tout simplement à la présence statistique d'un élément d'erreur spécifique dans une ou plusieurs des cinq études.

De plus, contrairement à la première étude de Wilson, cette analyse ne fournit pas de catégorie grammaticale pour les erreurs relevées, c'est-à-dire nous n'avons plus la distinction d’erreur de verbe, d'adverbe, de pronom et ainsi de suite. Les erreurs sont sélectionnées et présentées textuellement dans la mesure où le mot ou la phrase erronée est illustré(e) ou repris(e)

intégralement – à côté de laquelle on retrouve une indication de l'étude ou des études dans lesquelles l'erreur en question a été identifiée. Par exemple, nous avons les 69 « types d'erreur » distincts identifiés dans Wilson (1909) suivis d'une précision portant sur les différentes études dans lesquelles un même type donné a été signalé. Par la suite, les différents types d'erreurs identifiés dans les autres études sont présentés de la même manière contrastée. Cette liste comparative représente plus de la moitié de l'ensemble de l'article.

En ce qui concerne les résultats, trois des cinq études comparées ont obtenu des résultats similaires ou pour reprendre Wilson « almost identical results » : c'est-à-dire, les mêmes items lexicaux ont été jugés problématiques et ont été repérés dans trois études différentes. Cependant, les résultats des deux autres sont simplement présentés et les points de divergence ou de convergence ne sont ni examinés ni discutés en profondeur. Par contre, Wilson soutient que lorsque la liste des erreurs des élèves est soigneusement établie – chez les élèves-participants dès les premières années à l'école élémentaire – il y a très peu d'ajout de types différents dans les années supérieures (cf. « few errors are added by the upper-grade children »). Cela signifie simplement que les erreurs observées dans les premières années de l'enseignement primaire demeurent identiques d'un niveau à un autre et n'ont donc pas été corrigées de façon adéquate.

Ce constat traduit pour Wilson une sorte de fossilisation. En effet, malgré le maintien d'un certain nombre d'heures de cours de grammaire hebdomadaire tout au long du parcours scolaire des élèves, l'auteur voit dans ces répétitions la preuve que les erreurs ne seront pas uniquement le résultant d'un manque de compréhension ou de connaissances grammaticales, ou d'une règle de grammaire spécifique. Pour ce dernier, les classes de grammaire ne peuvent pas être en conséquence la solution. Il propose donc à nouveau ses exercices systématiques, qui selon lui ont déjà produit des « améliorations considérables » dans une école en particulier depuis sa première étude en 1909.

1922 - 1929 : Un pas vers une approche standardisée

Après avoir comparé les cinq études qui ont été inspirées par son modèle de 1909, Wilson constate beaucoup de différences méthodologiques chez les uns et chez les autres, notamment dans leur application - la façon dont les études ont été réalisées, la population étudiante ciblée, etc. - par rapport à son projet initial. Il a alors décidé d'élaborer un manuel ou une sorte de « mode d'emploi » de manière à réduire ces différences et favoriser une plus grande comparabilité entre les études futures. En 1922, il publie donc son premier guide d'étude qui a été révisé et augmenté en 1929.

Dans ces deux manuels, l'auteur a mis au point ce que l'on pourrait qualifier de test diagnostique d'erreur, avec des précisions quant à la date optimale de réalisation de l'étude, le raisonnement derrière sa nécessité d'implémentation (tant en recherche qu'en pédagogie), la manière dont le test doit être administré, entre autres. De plus, l'on y relève également des détails concernant ce que les "administrateurs" ou les enseignants ont le droit de dire aux élèves avant de faire passer le test. Malgré le fait que nous n'allons pas entrer dans une description détaillée de ce manuel, il est à noter que cette approche vise une plus grande standardisation de ces tests - ce qui en soi a de nombreux avantages notables (cf. chapitre IV pour notre positionnement sur ce point).

De manière générale, cette tendance traduit le besoin d'assurer, entre autres, une certaine reproductibilité et comparabilité entre études ayant un objet de recherche et une approche méthodologique similaires. Nous soulignons donc ici que ce besoin de rédiger un guide ou un manuel destiné à ceux qui effectuent des analyses d'erreur est mis en évidence d'une manière assez similaire dans la plupart des études et projets que nous examinons tout au long de cette thèse (cf. par exemple Corder 1967, les lignes directrices du projet de ICLE, etc.).

À titre d'information, le test dit d'erreur de Wilson était composé de six histoires courtes ; chacune comprenant 28 erreurs choisies en fonction de leur fréquence relative dans les études antérieures et de leur présence globale dans un « type d'erreur » donné (cf. l'étude de 1909). Le but de l'exercice était d'examiner la capacité des participants à reconnaitre les erreurs de langue les plus fréquentes, mais également de proposer des corrections. En ce qui concerne l'administration du test, il était conseillé que l'étude fût effectuée de préférence de manière longitudinale : à savoir à des intervalles de temps fixes - par exemple, en recommandant d'effectuer un nouveau test (une nouvelle histoire courte) toutes les six à huit semaines pendant une année entière ou au début, au milieu, et à la fin de l'année : les intervalles différentes visant à suivre l'évolution de la maîtrise de la langue (autocorrection ou de la reconnaissance d'erreur) chez les mêmes élèves.

En conclusion de cette section sur la contribution de Wilson dans l'avancement des études sur les erreurs en langue maternelle, ce qu'il est important de retenir dans ce contexte est la simplicité relative avec laquelle ces premières analyses ont été réalisées. Rappelons ici qu'il n'y avait pas de taxonomie d'erreurs ou d'hypothèses préconçues ou préétablies, mais qu'il s'agissait plutôt d'une approche empirique ascendante : l'on constituait un corpus d'échantillon de productions langagières réelles puis l'on procédait à une description de surface. Le résultat final était donc plus orienté vers la mise en place d'une liste d'erreurs et pas nécessairement une description ou une explication détaillée des causes ou sources de ces erreurs. On peut supposer alors que l'hypothèse

de travail était fondée uniquement sur l'identification des erreurs et qu'elle était considérée, par conséquent, comme suffisante à l'époque pour remédier, corriger ou empêcher toute reproduction des occurrences erronées préalablement identifiées.

1.2.1.2 Frei (1929)

Dans la même veine que les travaux de Wilson aux Etats-Unis au début des années 1900, l'Europe a également vu l'émergence d'un travail novateur portant sur les erreurs**,** à la suite de la publication de *la Grammaire des fautes* par Henri Frei en 1929. Ce travail a ouvert la voie à une plus grande discussion sur la place des erreurs, non pas dans une perspective pédagogique comme ce fut le cas ailleurs à l'époque, mais dans une perspective proprement linguistique en langue maternelle. Frei a en effet permis à la fois aux initiés et aux non-initiés de la recherche linguistique d'appréhender ce qui relève du « langage correct et incorrect » comme des véritables faits de langue dignes d'intérêt : et plus spécifiquement, l'auteur a démontré que l'incorrect ne renvoie pas systématiquement à des écarts aléatoires insignifiants.

A ce propos, ce dernier soutient :

> On ne fait pas des fautes pour le plaisir de faire des fautes. Leur apparition est déterminée, plus ou moins inconsciemment, par les fonctions qu'elles ont à remplir (plus grande expressivité, plus grande clarté, plus grande économie, etc.). Aussi est-il aisé de concevoir l'intérêt que présenterait pour le linguiste une étude fonctionnelle de ces faits (1929 :18)

Et en appuyant son raisonnement il rapporte la citation suivante du linguiste Bally :

> « Ajouterai-je qu'on devrait étudier systématiquement les incorrections ? Elles ont leur raison d'être, et répondent tantôt à des nécessités, tantôt […] aux exigences de l'expression émotive […] Dresser la liste de ces formes et les décrire, ce serait faire une besogne des plus utiles pour les linguistes à venir : si, parmi elles, les unes l'emportent, tandis que les autres restent sur le carreau, cela ne se fera pas sans raison… » (loc.cit)

C'est ainsi que Frei se donne pour objectif de « déterminer les fonctions que [les] fautes ont à satisfaire » dans l'usage linguistique. Ce faisant, il cherche à identifier tout ce qui ne relève pas d'une occurrence aléatoire et qui comporte de ce fait les traits d'une certaine systématicité – puisque la systématicité est envisagée comme l'expression de traits significatifs, inhérente à la langue et nécessitant donc une analyse approfondie. Il convient donc à notre sens de noter la contribution de ce travail remarquable – qui fut, à son époque, avant-gardiste tout en conservant encore aujourd'hui toute sa pertinence dans le contexte linguistique contemporain. Pour illustrer

tout le bien-fondé de ce travail, nous retraçons brièvement ci-après le contexte de son étude tout en mettant en avant la singularité de ses résultats.

Le caractère quelque peu inédit du travail de Frei l'a obligé à justifier son choix d'étude, non pas en termes de problématisation que l'on attribue généralement à la recherche aujourd'hui mais singulièrement par rapport à l'époque où la grammaire, en tant que science nouvelle, était envisagée sous un angle normatif, voire puritain. En effet, l'étude de la langue se devait à ce moment-là de se focaliser sur des formes normées et socialement acceptées, issues souvent d'un registre utilisé uniquement par des gens instruits. En guise donc de préface, le linguiste a jugé opportun de justifier sa démarche en adressant un message qu'il a soigneusement intitulé « à mon lecteur » - et dans lequel il maintient que l'écart systématique des erreurs de langage de toute entreprise scientifique peut être expliqué par leur nature jugée « sans intérêt », résultant de « la négligence ou [du] hasard », ou encore « comme des symptômes critiques annonçant le déclin de la langue ». Il s'est dit donc opposé à l'approche puritaine qui consiste à étudier ce qui doit être et non pas ce qui est.

Frei s'est donc constitué son corpus d'analyse principalement à partir de lettres échangées entre des prisonniers de guerre et leur famille. Il est à préciser que ces lettres « rédigées le plus souvent par des personnes de culture rudimentaire - généralement des femmes du peuple - expédiées de tous les coins de France » (ibid. : 28) étaient considérées comme reflétant « assez fidèlement l'état de la langue courante et populaire » de l'époque. (loc.cit) Les données ou échantillons du langage qu'il a utilisés proviennent de ce fait d'un corpus – tout à fait novateur par rapport à son époque et que l'auteur appelle « oral[8] », en raison du fait que les lettres ont été dictées par les prisonniers, puis retranscrites textuellement par une tierce personne. L'auteur s'est donc penché sur ce qu'il appelle « la vie des signes avec le seul souci de l'objectivité, pour rechercher en quoi les fautes sont conditionnées par le fonctionnement du langage et comment elles le reflètent » (ibid. : 8)

A la suite de son étude, son approche singulière a poussé Frei à concevoir l'erreur comme l'expression de ce qu'il appelle désormais « des déficits » de la langue. En effet, selon ce dernier la faute n'est pas forcément faute de la part du locuteur ; c'est en quelque sorte un symptôme de l'évolution de la langue, voire même le signe d'un dysfonctionnement du système linguistique. Il en déduit alors que les locuteurs en langue maternelle chercheraient inconsciemment à combler ces

[8] Notons plusieurs ambiguïtés potentielles dans l'usage et l'acceptation générale de ce terme : à savoir entre un discours spontané prononcé à l'oral, un autre pensé pour être écrit ou encore le fait que des lettres relèvent d'un genre textuel dit « écrit »

déficits, ce qui provoquerait des fautes systématiques. Les erreurs sont donc une manifestation de ceux-ci et renvoient à ce que Frei a catégorisé comme relevant des cinq principaux besoins linguistiques.

1) Besoin d'assimilation

Sont classées ici les erreurs relevant d'une forme d'analogie ; l'on confond le sens (analogie sémantique) ou la forme correcte d'un mot (analogie formelle). Les locuteurs procéderont alors soit en attribuant une nouvelle interprétation à un item lexical ayant une certaine similarité avec un autre, soit en « créant » une nouvelle forme à partir d'une construction existante. L'idée pour le sujet parlant est de ramener l'inconnu au connu, de remotiver ou de réactualiser quelque chose de déroutant.

Ex : Les adjectifs « indifférent » et « stupéfait » sont interprétés comme des participes (présent ou passé) et permettent la création de **cela m'indiffère* ou **j'ai été stupéfait par cette nouvelle*.

Ex : L'expression *découvrir le pot au rose* est souvent orthographiée de manière erronée : « le poteau rose » ou le « pot aux roses ». Le terme renvoyait à l'origine au pot de maquillage avec lequel se fardaient les acteurs (d'où la notion de découvrir le pot au rose, c'est-à-dire découvrir ce qui se cache sous le fard), puis il y a eu réactualisation sémantique.

Ex : *Avoir affaire* est réactualisé en **avoir à faire*, et ainsi de suite.

2) Besoin de différenciation

Les erreurs commises dans le but d'éviter les confusions « latentes ou réelles » (ibid. : 62) de type équivoque, polysémie ou homophone relèvent ici des procédés de différenciation.

Ex : Pour montrer un cas d'équivocité, **c'est lui quila fait venir* peut être interprété de quatre façons différentes augmentant de ce fait la fréquence d'emploi erronée (cf. qu'il a ? qui la ? qui l'a ? …)

3) Besoin de brièveté

Comme son nom l'indique, le principe opératoire ici est de faire économie par brièveté à travers des procédés fournis par le système linguistique – notamment, entre autres par la troncation, l'apocope (proprio, mémo, apéro) ou l'aphérèse (car, bus pour autocar ou autobus, etc.)

4) Besoin d'invariabilité

Les erreurs relevées ici se portent aussi bien sur les items lexicaux que sur les items grammaticaux. Elles renvoient à la difficulté de choisir l'usage correct entre une pléthore de possibilités, tandis qu'un mot ou forme « fourre-tout » et invariable suffirait.

Ex : « La langue écrite, sous l'action du besoin de clarté, tend à exprimer les diverses corrélations au moyen de procédés explicites : *parce que, puisque, pour que, pendant que, au point que, sans que,* etc. La tendance populaire, au contraire, est de remplacer tous ces signes par un instrument unique — le corrélatif générique *que* » (ibid. 153)

5) Besoin d'expressivité

   Ce besoin, qui ne doit être compris qu'en tant qu'esquisse ou « première observation » dans les termes de Frei, renvoie à un ensemble de phénomènes hétérogènes par leur nature et dans lesquels les locuteurs chercheraient à exprimer une certaine « affectivité ». Cela étant, ce besoin se distingue des quatre précédents dans la mesure où ils répondaient à un véritable besoin ou souci de communication tandis que ce dernier se place plutôt du côté des caractéristiques individuelles. Ces traits ne sont pas sans rappeler les erreurs dites idiosyncratiques qui ont été largement théorisées par Corder (cf. section 1.2.2.2).

Nous pouvons donc résumer cette section en soutenant que le travail de Frei a constitué - au moment de la publication - une percée importante dans la linguistique au vu notamment de son approche et sa description systématiques des phénomènes relevés, ainsi que le choix osé de son corpus « oral ». On peut également le « féliciter » par la suite d'avoir permis à l'analyse des erreurs d'être propulsée au premier plan des études linguistiques, et ce, dès l'année 1929. Notons toutefois que sa longue liste d'erreurs est plusieurs fois tombée dans la même approche puritaine que celle à laquelle il s'était farouchement opposé au début de son livre. Ceci a eu pour résultat quelques généralisations abusives et quelques descriptions prescriptives, notamment dans sa comparaison du français parisien avec les autres variétés examinées. En somme, nous reconnaissons que son travail peut être considéré pionnier dans la linguistique française et européenne et d'avant-garde, avec de nombreux apports qui gardent toute leur pertinence aujourd'hui. Mais la transposition de ses concepts en langue étrangère s'avère difficile, dans la mesure où l'erreur chez Frei répond tout d'abord à un besoin sous-jacent de la langue elle-même et ne renvoie vraisemblablement pas aux exigences de la langue qui n'auraient pas été maîtrisées chez le locuteur.

1.2.2 En langue étrangère

Comme nous l'avons vu dans la section 1.2.1, il est possible d'examiner les erreurs d'un point de vue dit linguistique tout en adoptant plusieurs angles d'approches différentes : à savoir, d'un point de vue purement pédagogique comme dans le cas de Wilson ou proprement théorique (et ou descriptif) dans le cas de Frei. En effet, ce constat est valable pour les études portant à la fois sur les erreurs commises en langue maternelle et en langue étrangère. Cela étant dit, nous allons maintenant poursuivre l'examen des travaux de trois linguistes supplémentaires, mais cette fois-ci vis-à-vis de l'erreur en langue étrangère. Ces travaux, que nous considérons comme fondateurs, ont profondément contribué à la recherche et l'enseignement des langues étrangères et ils ont tout particulièrement permis d'apporter un regard avisé sur le repérage, la description et l'exploitation des erreurs tant pour le chercheur que pour l'enseignant sur le terrain.

De plus, en dépit des nombreux travaux qui leur ont succédé, les études des trois auteurs que nous présentons dans les sous-sections ci-après demeurent très pertinentes dans le contexte linguistique actuel. Par exemple, dans le cas des travaux de Palmer (1917, 1922 …), son travail de pionnier sur l'importance du vocabulaire spécialisé (cf. The General Word List), les collocations et bien d'autres pistes de recherche lexicographique continuent à être explorées et exploitées aujourd'hui. La désormais célèbre taxonomie d'erreurs de Corder (1967, 1973) a jeté les bases de ce que l'on appelle de nos jours l'analyse d'erreurs – et malgré quelques critiques modernes, tous ses travaux servent encore de référence pour ceux qui souhaitent explorer les erreurs en didactique des langues étrangères. Enfin la contribution de James (1998) se présente comme un examen très exhaustif - en fournissant une cadre de réflexion à la fois factuel et critique vis-à-vis de l'analyse des erreurs dans la recherche et l'enseignement des langues étrangères.

1.2.2.1 Palmer (1917, 1921, 1924, …)

Comme nous venons de le mentionner ci-dessus, l'appréhension des erreurs en tant qu'objet de réflexion ou de recherche n'est pas restreinte aux études entreprises « en » et « sur » les langues maternelles. De plus, les débuts conceptuels de l'avènement de ces études en langue étrangère ne seraient pas aussi récents que certains le prétendent. En effet, il nous semble important de souligner qu'Harold E. Palmer, considéré comme un des pères fondateurs de la linguistique appliquée telle que nous la concevons aujourd'hui – et tout singulièrement dans l'imaginaire historique britannique (cf. Smith 2011) – a vu dans les erreurs un objet nécessitant une réflexion critique voire des études approfondies : et ce, dès 1917.

Comme certains qui l'ont précédé ou lui ont succédé, c'est suite à son expérience en tant qu'enseignant d'anglais langue étrangère que Palmer a décidé de se mettre à la « théorisation » de la pratique pédagogique en elle-même – notamment à travers des écrits qui visaient un lectorat d'enseignants comme lui-même ou des décideurs des politiques linguistiques institutionnelles[9]. Son travail a été rapidement assimilé et vulgarisé (notamment au Japon où il fut enseignant) avant d'être intégré au courant linguistique européen de l'époque. Aujourd'hui des travaux notables qui lui sont attribués ont contribué au développement de domaines de recherche qui sont encore d'actualité aujourd'hui. Par exemple, il a été l'un des premiers à publier sur les sujets suivants : l'importance de l'étude de la collocation (dans le but de réduire la marge d'erreurs des apprenants), des listes de

[9] Notons à titre d'information qu'il fut l'expert en didactique de langue anglaise auprès du gouvernement japonais dans les années 1920. A ce titre, il a joué un rôle considérable dans le développement de certaines méthodes d'enseignement (cf. Smith 1999)

vocabulaire (cf. the General Service List of English Words)[10], des dictionnaires d'apprentissage, et ainsi de suite.

Toutefois, bien que Palmer n'ait pas clairement identifié l'analyse des erreurs comme une entreprise indépendante à part entière, il convient de noter qu'il a grandement mis en avant le besoin d'avoir des cours de langue étrangère qui devaient être de « nature corrective » - ce qui suggère une certaine propension à concevoir les erreurs comme une partie intégrante du processus d'apprentissage d'une langue et cela d'autant plus que les premières tentatives ou productions langagières seront ostensiblement déficientes. Il en va de même dans le raisonnement de Palmer qui estime que cela nécessiterait une attention particulière, au début dès l'apparition systématique des erreurs jusqu'à leur « extinction quasi-totale ». De même, il incombait à l'enseignant d'établir une approche corrective afin d'anticiper et remédier toute apparition d'occurrences erronées. Palmer ne parle pas alors de « faute » ou « d'erreur » ; mais il se réfère à ces événements en termes de « déficit », de « l'inexactitude », de « mauvaises habitudes », « pidgin-speech », voire même de « tendances vicieuses » que les professeurs de langue doivent expliquer consciencieusement aux étudiants – afin de pouvoir les corriger de manière optimale.

Sans trop nous attarder sur d'autres précisions concernant les erreurs ou l'évaluation de la production langagière en langue étrangère – qui ont, précisons-le, bénéficié de presque cent ans de recherche depuis le travail novateur de Palmer – soulignons que ce dernier soutenait l'idée que les erreurs s'expliquaient de deux façons.

Soit l'erreur était synonyme d'une approche défaillante dans la méthode d'enseignement :

> We would urge that the factor of error should never be allowed to obtain any footing at all. All errors other than those made by native speakers are abnormalities, and the result of a faulty method (1917 : 119)

Soit ce qui était demandé n'était pas approprié ou adapté aux besoins et compétences réelles de l'apprenant :

> If the work of a serious student is characterized by a certain proportion of error, it is fairly a trustworthy sign that he is doing work which is too difficult for him [...] All work performed by students in accordance with a properly graduated method under ideal conditions should be marked by extreme facility and extreme accuracy. (ibid.: 121)

[10] Ce travail fait suite à des recherches effectuées par Palmer - et a en bénéficié en grande partie -, notamment les premières listes de mots destinés à un public d'apprenants spécifiques. (cf. Palmer 1933 ; Smith 1999, 2011).

Cela étant dit, en dépit du fait que nous ne soutenons pas le positionnement adopté par ce dernier en termes de description, d'explication des sources d'erreurs, et de l'attitude à adopter vis-à-vis de celles-ci, il est à souligner que les travaux de Palmer ont été largement repris et poursuivis de manière générale par ses successeurs. Cependant, il a tout de même fallu plus de cinquante ans pour que les erreurs de langue étrangère redeviennent une question actuelle et surtout pour qu'on leur accorde une méthodologie propre.

1.2.2.2 Corder (1967, 1971*a*, 1971*b*, ...)

Les années 1960 et 1970 voient un tournant dans le regard porté sur les erreurs par les linguistes, redonnant ainsi un nouveau souffle aux études d'erreurs qui jusque-là étaient effectuées par quelques chercheurs isolés. En effet, ce n'est vraisemblablement qu'à partir des premières publications de Stephen. P. Corder que « l'analyse des erreurs », telle qu'on la connait aujourd'hui, a commencé à s'imposer comme un véritable objet de recherche nécessitant plus que des critiques des enseignants et un regard furtif des linguistes. A notre sens, Corder a été en quelque sorte l'initiateur de ce regain d'intérêt, en apportant notamment une explication critique eu égard à la nécessité d'analyser les erreurs d'un point de vue linguistique. Il a, à ce titre, proposé une méthodologie propre.

En effet, Corder (1971b) affirme que l'étude des erreurs constitue un double enjeu, aussi bien pour le linguiste théoricien que pour l'enseignant. Il s'agit tout d'abord, selon lui, d'élucider « lorsqu'un apprenant étudie une langue étrangère, ce qu'il apprend et comment il apprend ; ce but est d'ordre théorique » pour comprendre ensuite « le but appliqué » *:* à savoir « comment [les erreurs] sont produites et pourquoi » (p.26)[11]. Pour ce dernier, la motivation principale derrière ces études est le besoin de mieux rendre compte du phénomène en termes d'erreurs du système langagier qu'aurait l'apprenant à un moment donné. Ceci fait écho à la logique d'Henri Frei (1929) qui préconisait que l'on mette systématiquement les erreurs en relation avec le système qui les a produites, puisqu'elles sont d'un côté les symptômes de l'évolution de la langue tout en étant les produits normaux de toute acquisition linguistique.

Cependant, il semble judicieux de nous en tenir ici à Corder, tout en rappelant le positionnement quelque peu normatif et singulièrement diachronique de Frei : rappelons que ce dernier voyait dans la faute un moyen de réparer les déficits du langage correct dans la langue maternelle et ses

[11] Cf. Corder 1980b pour la version française de l'article de Corder 1971b, publié initialement en anglais.

descriptions des fautes en termes de fonctionnement du langage sont foncièrement d'un point de vue évolutif. A l'inverse, Corder soutient que son positionnement est à mettre en relation avec le déclin des théories d'analyses contrastives[12] de l'époque (cf. Robert Lado 1957, Granger 1996) ou en réponse à tous ceux qui soutiennent que les erreurs *développementales*, *interlinguales*, ou *idiosyncrasiques* indiquent que la langue-cible n'est pas encore acquise et qu'il faut une pratique intensive des formes correctes pour y remédier.

L'objectif donc de l'analyse des erreurs de Corder (1980b) est double. D'une part, à destination de l'enseignant pour qu'il ait un outil, dans la mesure où celui-ci permet de relever et de mieux expliquer les hypothèses qui ont provoqué les erreurs, et, d'autre part, à destination de l'apprenant pour qu'il puisse avoir « des informations ou des données utiles pour une meilleure compréhension de telle ou telle règle de la langue-cible » (ibid. : 27) par rapport aux erreurs commises. Ainsi, avec cette méthode, Corder pense éviter que l'enseignant écarte hâtivement les erreurs lors de l'évaluation/correction, sous prétexte qu'elles ne seraient d'aucune importance ou encore « as possible annoying, distracting but inevitable by-products of learning a language » (1967 : 162).

Passons maintenant à la méthode d'analyse préconisée par Corder (1967, 1973), qui s'établit en trois étapes, à savoir la reconnaissance, la description et l'explication des erreurs. Une méthode certes prenante pour l'analyste qui doit établir avant tout une liste exhaustive des erreurs issues de la production écrite (ou orale) de l'apprenant, tout type confondu, pour en faire son corpus d'étude. Il s'agit tout d'abord ici de la RECONNAISSANCE. Vient ensuite la première différenciation entre les éléments recueillis, selon la taxonomie de Corder, selon qu'ils sont considérés comme étant des « erreurs, lapsus linguae ou fautes[13] » : ou encore des erreurs dites « systématiques » ou « non-systématiques ».

Sont considérées comme systématiques, les erreurs dites de compétence (au sens chomskyen) et qui « reflètent à un moment donné [la] connaissance sous-jacente, ou, comme on pourrait l'appeler, [la] compétence transitoire » des apprenants en langue étrangère (Corder 1980a : 13). Les lapsus et fautes, envisagés comme des erreurs de performance et donc non-systématiques, relèvent « *des aléas accidentels dans la performance linguistique et ne reflètent pas les lacunes de la connaissance de la langue* » (ibid. : 13). Soulignons ici que Corder considère que les erreurs non-systématiques (ici, les fautes et les lapsus) ne sont pas significatives par rapport au processus d'apprentissage d'une langue, et qu'elles sont commises tant par l'apprenant de langue étrangère

[12] Cf. section 1.3.2 pour une présentation brève.
[13] Ces différents termes seront explicités en section 1.3.1

que le natif. Il est question donc lors de cette étape dite de RECONNAISSANCE de relever toutes les erreurs dans le but de procéder à une première tentative de catégorisation, selon que ces dernières s'avèrent systématiques et donc analysables ou non-systématiques et de ce fait négligeables.

Notons toutefois que la notion de faute chez Corder semble renvoyer aux erreurs dues au hasard des circonstances, telles des défaillances de mémoire, la fatigue, etc. Ces circonstances s'avèrent à notre sens inidentifiables et subjectives, sans l'intervention explicite de l'apprenant-scripteur. Notons, de plus, une ambiguïté terminologique au niveau de ces concepts clés préconisés par Corder. Certains de ses travaux parlent de « fautes, lapsus et erreurs » tandis que d'autres semblent désigner ces mêmes notions sous le terme d'erreurs dites « présystématiques », où l'apprenant n'est que [vaguement] conscient d'une règle à appliquer dans une situation donnée en L2, systématiques et post-systématiques (lapsus). » (cf. notamment 1967 et 1973 respectivement). Examinons brièvement le tableau 2 suivant où Corder nous livre une première tentative de catégorisation, envisagé certes de manière rudimentaire sans pour autant en être simpliste, afin d'aider l'analyste à identifier les erreurs porteuses d'information et sur lesquelles on peut procéder à une explication linguistique approfondie. Il en ressort que les deux autres étapes de l'AE de l'auteur s'intéressent uniquement aux erreurs du type 2 et 3.

Type d'erreurs	Correction possible[14]	Explication possible [15]
i. Présystématique	Non	Non
ii. Systématique	Non	Oui
iii. Post-systématique	Oui	Oui

Tableau 2 : Classement d'erreurs proposé par Corder 1973

La deuxième phase est celle de la DESCRIPTION. La catégorisation entamée est poursuivie dans le but d'en faire ce qui s'apparente à un classement linguistique, par opposition au classement purement typologique/pragmatique pour ce qui est de la première. Les erreurs sont à nouveau reconfigurées et se voient attribuées à ce que Corder appelle les « niveaux linguistiques différents ». Ici, on procède par comparaison de la phrase erronée avec sa reconstruction dans la langue cible pour ensuite classer les erreurs selon qu'elles relèvent d'un principe d'omission, addition, sélection ou d'ordre des mots *:* l'omission étant l'oubli d'un élément nécessaire (pour la construction de la phrase) ; l'addition l'ajout d'un élément inutile ou incorrect ; la sélection d'un élément incorrect (en termes de choix lexicaux) ; ou un problème d'ordre des mots.

[14] De la part de l'apprenant
[15] Du linguiste/enseignant

Ces quatre niveaux ne pouvant permettre une description / catégorisation suffisante des erreurs, Corder (1973) préconise un sous-classement non exhaustif dans lequel les quatre types peuvent être expliqués, par exemple selon un ordre phonologique, grammatical, etc. Nous avons reproduit ci-après un tableau illustrant la portée de cette classification.

	Phonologique / Orthographique	Grammatical	Lexical
Omission			
Addition			
Sélection			
L'ordre des mots			

Tableau 3 : Adaptation de la matrice de classification des erreurs

Malgré l'apparente opérationnalité de cette classification, il est important de souligner qu'elle s'avère réductrice et inefficace dans la mesure où elle permet le classement d'une même erreur dans des cases différentes, ce qui pourrait provoquer d'importantes contradictions. De plus, ces contradictions peuvent aboutir à des généralisations massives où l'erreur se voit hâtivement classée de façon approximative. On remarque également que ces catégorisations demeurent strictement *de surface :* autrement dit, elles ne nous renseignent pas sur le « *pourquoi* » des erreurs ou le système sous-jacent derrière leurs apparitions ou encore moins ne nous expliquent comment les corriger, ce qui est d'ailleurs l'objectif final de l'AE.

Ces lacunes et éléments restés jusque-là inexpliqués motivent la troisième et dernière étape que Corder nomme ingénieusement l'EXPLICATION. En effet, Corder rappelle qu'elle « est principalement psycholinguistique dans la mesure où on essaie de voir pourquoi le dialecte idiosyncrasique[16] de l'apprenant est comme il est, et comment il se construit » (1980b : 26). Donc on passe d'une étape où on identifie les erreurs comme « analysables » ou non, pour ensuite dresser une typologie principalement descriptive dans le but de mieux expliquer ce qui provoque les erreurs, en partant bien entendu de l'hypothèse qu'elles proviennent (toutes) du dialecte idiosyncrasique de l'apprenant.

D'autre part, Corder souligne qu'une autre explication qui ne renvoie aucunement à la langue source de l'apprenant, demeure plausible. Il s'agit en effet du *principe d'analogie*, autrement dit si l'on considère l'apprentissage d'une langue étrangère comme étant un processus principalement cognitif, les erreurs peuvent être provoquées par de « fausses hypothèses » qui seraient basées sur le système langagier de la langue cible que l'apprenant aurait intériorisé à un stade donné. Un

[16] Terme utilisé par Corder pour désigner ce que l'on appelle communément « l'*interlangue* » de l'apprenant. Il est à noter que nous choisissons de ne pas reprendre cette terminologie dans le présent travail pour éviter tout amalgame avec d'autres termes renvoyant au même référent. Cf. section 13.2 pour une présentation détaillée.

système qui serait donc incomplet. L'apprenant chercherait dans ce cas précis à transposer des « règles réellement apprises » à des situations incompatibles avec celles-ci. Ce qu'il faut retenir alors de cette explication c'est l'idée de la *surgénéralisation* d'une règle de grammaire que l'apprenant penserait maîtriser, ce qui provoque ce que certains didacticiens appelleraient « une bonne faute ».

En définitive, malgré l'avancée de Corder qui préconisait que l'on incorpore les erreurs faites par l'apprenant dans une analyse linguistique approfondie, la méthode ci-dessus décrite n'est pas sans défaut. Des défauts que nous jugeons non-négligeables. D'une part, nous rejoignons l'argument de Perdue (1980) qui soulignait l'aspect limitatif de ladite méthode. Premièrement, la matrice classificatoire susmentionnée permet de classer une erreur de plusieurs façons « concurrentes », ce qui pose d'emblée un problème méthodologique. Deuxièmement la distinction entre erreurs *systématiques* ou *non-systématiques, erreurs/fautes, analysables ou négligeables* s'avère problématique dans la mesure où Corder lui-même affirme qu'elle repose sur une interprétation subjective de la part de l'analyste, notamment si ce dernier n'a pas recours à l'apprenant pour vérifier l'exactitude de l'énoncé. Autrement dit pour vérifier l'intention communicative[17] du message.

D'autre part, la taxonomie linguistique privilégiée ici laisse entendre qu'elle s'applique principalement à des erreurs faites par des apprenants débutants ou intermédiaires, puisque les erreurs relevées et décrites sont principalement des erreurs dites locales. De plus, l'analyse s'opère à un niveau principalement « phrastique » où l'erreur est ramenée à la phrase dans laquelle elle est produite sans prendre en compte le contexte « textuel » ou moins encore « *discursif* ». De ce fait, bien que ces niveaux ne correspondent que dans une moindre mesure à notre corpus d'étude, nous soutenons que l'AE, telle que présentée par Corder, n'est pas tout à fait compatible avec plusieurs objectifs sous-jacents de la présente étude – notamment en raison du fait qu'elle ne permet pas d'expliquer de manière suffisante ce qui provoque les dysfonctionnements au niveau textuel. Il nous faut donc trouver un cadre complémentaire qui « comblerait les déficits ».

[17] Selon Corder (1980b), dans les énoncés qui se veulent difficiles à comprendre, il faut, dans la mesure du possible, demander à l'apprenant d'expliquer le but de son message (éventuellement dans sa langue maternelle) – sans quoi nous n'aurions qu'une reconstruction plausible, sans en être certains de la véritable intention du message initial. Ce qui pose bien entendu des problèmes méthodologiques que nous ne pouvons pas développer ici, faute de temps.

1.2.2.3 James (1998)

Contrairement aux auteurs présentés dans les sections 1.2.1.1 jusqu'à la section 1.2.2.2, celui qui suit ne s'appuie pas sur des éléments de corpus dans son exploration des erreurs. Il se positionne dans son ouvrage de façon principalement descriptive voire proprement théorique. Et à cet égard, il constitue un des premiers à se consacrer entièrement aux erreurs en langue étrangère – qui plus est, dans une perspective proprement linguistique. En effet, dans celui-ci son auteur, Carl James, y dépeint la place qu'occupe l'analyse des erreurs en linguistique appliquée, en linguistique générale et même son positionnement face à ce qu'il appelle les différentes « théories linguistiques ».

Ce dernier postule que « the study of human-error making in the domain of language error analysis is a major component of core linguistics » (1998 : 93), tout en soulignant que l'analyse des erreurs ne suppose pas une « branch of linguistic theory (or pure linguistics) but of applied ». Il est important de ne pas se méprendre sur la portée de cette affirmation dans laquelle James explique que l'AE n'est pas à mettre au même niveau que des courants théoriques, tels que le générativisme ou le fonctionnalisme mais qu'elle renvoie à une approche plus concrète. Notamment dans la mesure où elle est directement applicable et exploitable sur le terrain. A ce titre, l'approche ici se veut ascendante et suppose, de ce fait, une certaine propension à recueillir et à étudier des données réelles dans le but, dans un premier temps, d'en exploiter les éléments les plus significatifs pour ensuite procéder à leur description. Ceci est à mettre en opposition avec certaines théories linguistiques qui ne prennent pas les données comme point de départ, mais au contraire comme moyens de vérifier des hypothèses ou des abstractions déjà établies.

Sans tomber dans le sensationnalisme, l'auteur y dresse un portrait objectif de l'analyse des erreurs à la fois (i) en retraçant les critiques qu'a rencontrées le courant dès son émergence jusqu'à nos jours et (ii) en établissant un état des lieux des tenants majeurs qui ont avancé une sorte de raison d'être du courant de l'AE en linguistique, sans oublier bien entendu ceux qui l'ont, d'une manière ou d'une autre, incorporé à leurs propres recherches. A titre d'illustration, l'auteur rappelle que l'analyse des erreurs (telle que mise en avant par Corder) fut considérée comme la remplaçante légitime du courant précédant appelé « analyse contrastive » dans lequel les erreurs en langue étrangère furent explicitées principalement en termes d'interférence avec la langue maternelle. Ce paradigme de comparaison accrue fut de courte durée, étant donné qu'il ne permettait pas de rendre compte de nombreux phénomènes observés par ceux qui s'intéressaient aux erreurs en langue étrangère comme la manifestation évidente d'un transfert du système de la langue maternelle.

A travers le livre de James qui se lit comme un recueil ou une grande synthèse, des travaux antérieurs sont réactualisés par le biais des commentaires d'un linguiste avisé et non seulement d'un enseignant de terrain qui observe des phénomènes de fil en aiguille (ce qui n'enlève, à notre sens, ni le mérite ni la recherche de l'objectivité). En effet beaucoup de ceux qui se sont intéressés à l'analyse des erreurs l'ont fait d'un point de vue foncièrement pédagogique en ce que le point de départ fut l'enseignant et ses élèves, comme nous l'avons mentionné dans les sections précédentes – et qui, de plus, demeure bien entendu compréhensible au vu de la conjoncture. Le travail de James se veut toutefois rassembleur d'un point de vue strictement linguistique dans la mesure où l'auteur s'est donné comme objectif de « dispel any remaining misconceptions about EA being a narrow academic pursuite » (ibid. 25). En définitive, l'ouvrage de James a permis de consolider l'approche montante de l'analyse d'erreur, notamment en soulignant que « the scope of EA is wide and widening » (loc. cit).

1.3 Quelques définitions et taxonomies résultantes

A travers le bref rappel historique ci-dessus, nous avons illustré le fait que l'étude de l'erreur n'a pas une origine homogène. Il serait donc erroné de dire que l'AE tel que nous le concevons aujourd'hui est le résultat d'un travail précurseur d'une seule personne ou même le résultat d'un groupe de chercheurs spécifiques. De plus, il est également important de reconnaitre d'une part, que les cinq principaux auteurs cités en section 1.2 ont été choisis par rapport aux approches et tout particulièrement leurs débuts conceptuels très différents les uns des autres et d'autre part, que d'autres auteurs ont également apporté par la suite leur lot de contributions au cadre de l'analyse des erreurs à travers de multiples publications.

Les nombreux « précurseurs » et notamment ceux qui leur ont succédé ont apporté des contributions spécifiques en se positionnant soit en accord direct soit en divergence avec un modèle existant. Ce faisant, l'appréhension de ce qui doit être considéré comme une erreur est en partie adoptée et approfondie ou, dans certains cas, remplacée ou complètement écartée au profit de nouvelles terminologies. Par exemple, ceux qui adoptent l'approche de Frei pourraient considérer les erreurs principalement sous un aspect moins stigmatisant en termes d'évolution naturelle du langage dont le but intrinsèque (des erreurs) serait de compenser certaines lacunes inhérentes à la langue ; tandis que ceux qui se positionnent dans l'approche de Corder pourraient voir dans les erreurs des occurrences systématiques ou des occurrences non-systématiques (lapsus, fautes) – conduisant de ce fait à des analyses totalement différentes, sans une once de complémentarité évidente.

Comme conséquence directe de l'apparition et de l'évolution des nombreuses théories linguistiques au fil des publications, la notion d'erreur, sa portée voire l'explication de ses origines chez l'apprenant ont donné naissance à de nombreuses terminologies différentes. Ces termes clés issus des divers travaux d'analyse d'erreur sont utilisés avec un certain degré de fréquence dans la littérature actuelle, mais aussi avec un certain degré de dissemblance dans leur sémantique. Examinons ci-après quelques-uns de ces termes clés et notre positionnement quant à ces différences de sens multiples.

1.3.1 Erreur, faute et écart

Depuis l'avènement de l'analyse des erreurs sous la forme que l'on connait aujourd'hui, à l'instar notamment de Corder (1967, 1971a, 1971b) et Perdue (1980) beaucoup ont suggéré que les erreurs, fautes ou écarts devraient systématiquement être différenciés lors de l'évaluation de toute production erronée - que ce soit en langue maternelle ou en langue étrangère. Cela nous amène à nous demander pourquoi et comment les définir ou les différencier, et surtout si cette différenciation est utile dans la présente étude. Tout d'abord, le raisonnement derrière cette distinction réside dans le fait que l'utilisation incorrecte de la langue peut être due à plusieurs facteurs : à savoir le manque de connaissances adéquates (donc l'ignorance d'une norme ou d'une règle spécifique) ; l'utilisation erronée par inadvertance où le contexte situationnel joue un rôle important (l'état psychologique du locuteur, fatigue ou autre) ; ou les occurrences non systématiques et aléatoires qui se produisent sans que l'on puisse les classer comme appartenant au premier et deuxième facteur établis ci-dessus.

Toutefois, selon que l'on s'identifie à l'enseignant en classe de langue ou au chercheur en linguistique - les perspectives peuvent être diamétralement opposées, tout en étant complémentaires. L'enseignant pourrait, d'une part, vouloir faire la différence entre les erreurs de manière à rendre compte de la progression de ses élèves dans la mesure où l'usage incorrect d'un item spécifique pourrait nécessiter différents types de remédiation ou correction pédagogique. D'autre part, le chercheur pourrait ne pas être intéressé par l'incidence pédagogique immédiate des erreurs de langue, mais chercherait plutôt à mieux comprendre le processus qui sous-tend ou conduit aux dites erreurs, et par conséquent à mieux décrire et distinguer différents types d'usage erroné[18].

[18] Ce faisant, le linguiste pourrait chercher à créer par exemple une taxonomie d'erreur. Cf. Marquilló Larruy (2003) pour un exemple d'usage ou Andersen (2011) pour un bref aperçu de quelques taxonomies existantes.

Nous notons cependant que malgré la diversité des acteurs s'intéressant à l'analyse des erreurs, les définitions les plus souvent avancées sont celles qui renvoient dans le moindre détail à celle mise au point par Corder. Ce besoin incessant de distinction se voit mis en avant dans bien des études, soit pour mettre en cause les premières définitions en y ajoutant de nouveaux traits distincts (cf. James 1998 ; Marquilló Larruy 2003) - soit tout simplement pour réactualiser une définition jugée « ancienne ». A notre sens, la reprise de la définition de Corder peut s'expliquer légitimement de plusieurs manières. Nous en avancerons deux : (i) soit par rapport à la concision et la clarté qu'elle englobe ; (ii) soit parce qu'il n'y a pas de marge d'interprétation heuristique ou normative. Cela dit, il nous semble judicieux de la reproduire ici :

> Les erreurs de performance seront par définition non -systématiques, et les erreurs de compétence systématiques. […] Aussi sera-t-il commode désormais d'appeler « fautes » les erreurs de performance, en réservant le terme d'« erreur » aux erreurs systématiques des apprenants, celles qui nous permettent de reconstruire leur connaissance temporaire de la langue, c'est-à-dire leur compétence transitoire. (1980a : 13)

Outre la dichotomie entre erreur et faute, on retrouve d'autres notions périphériques : par exemple la notion de *lapsus linguae* chez Corder, *déviance* employée comme terme générique dans James (1998), ou encore *écart* utilisé par certains comme un fourre-tout. Toutefois ces terminologies mènent souvent à davantage d'ambiguïtés plutôt que d'apporter les éclaircissements visés. Prenons le cas du lapsus linguae qui renvoie aux erreurs dues au hasard des circonstances, telles que les défaillances temporaires de mémoire, la fatigue, etc. Ces circonstances demeurent souvent subjectives et invérifiables sans l'intervention formelle de l'apprenant-scripteur, qui demeure le seul à même de préciser son intention communicative – c'est-à-dire, si tel ou tel emploi résulte de l'inadvertance.

Bien que nous accordions une certaine utilité à ces distinctions dans les analyses d'erreur à petite échelle, à savoir en classe de langue ou avec un corpus restreint, celles-ci n'ont aucune incidence directe dans notre étude. Et ce, dans la mesure où toutes les occurrences jugées erronées seront relevées : qu'elles soient donc des fautes, erreurs ou autre. Il nous reviendrait alors de procéder à une analyse qualitative sur une occurrence erronée identifiée en tant que telle par rapport à sa fréquence dans l'ensemble de notre corpus d'étude. A titre d'exemple, une « faute » dans la terminologie de Corder serait non-systématique et devrait par conséquent se traduire par une fréquence moindre dans le corpus : ceci est le cas pour certains items annotés en tant qu'erreurs orthographiques, étant donné leur caractère non-systématique ou aléatoire. Mais comme nous

verrons dans le chapitre VI, ce n'est pas toujours le cas et la définition avancée est mise à rude épreuve face aux résultats statistiques.

1.3.2 Interférence, interlangue et transfert

Ces trois termes sont à mettre en rapport avec le cadre d'analyse qui a précédé l'AE, à savoir l'analyse contrastive. Dans sa forme la plus appauvrie, l'hypothèse ici était de soutenir que l'acquisition d'une langue étrangère et conséquemment son utilisation dépendraient de manière considérable de la structure de la langue source, et donc que les erreurs des apprenants étaient prévisibles et interprétables en fonction de cette dernière. Il s'agissait alors de procéder par analyse des convergences et des divergences notables dans deux systèmes linguistiques, l'objectif étant d'identifier les points communs qui pouvaient se prêter à une acquisition « facile » en L2 et d'autres qui s'avéreraient plus contraignants. Surgit ici alors le concept de transfert de « forme et sens » aussi bien en réception qu'en production en langue étrangère.

De manière succincte, l'analyse contrastive privilégiait la langue maternelle comme principale source d'erreurs dans la mesure où celle-ci était censée provoquer ce que certains désignent comme étant un problème de *transfert* (Richards 1974) ou d'*interférence* (Selinker 1972). Ces deux termes peuvent prêter ordinairement à confusion étant donné qu'ils renvoient tous les deux à un phénomène quasi-identique que Corder lui-même ne distingue pas : à savoir « the inappropriate use of the rules of [the] mother tongue in […] performance of the target language » (1973 : 132). Ce qui permet de les distinguer sera donc l'acceptabilité de l'énoncé : le premier ne présentant pas d'erreur jugée inacceptable (transfert positif) tandis que le deuxième constitue une erreur dans la langue cible (transfert négatif).

Quant à la notion d'interlangue, cela renvoie à une compétence transitoire qu'aurait l'apprenant au cours de l'apprentissage d'une langue étrangère (cf. Selinker, 1972 ; Hasselgard & Johansson, 2011). A ce stade, il n'a pas, à proprement parler, une compétence parfaite dans la langue cible, mais aurait déjà intériorisé une grammaire approximative de celle-ci. Signalons ici que l'existence où l'étendue de cette nouvelle langue ou grammaire en construction ne peut être étudiée qu'avec parcimonie à travers des tests de « compétence », en termes notamment d'élicitation explicite des règles apprises.

1.3.3 Tendance actuelle : vers une approche textuelle de l'erreur

Comme nous venons de mentionner, les différentes définitions de l'erreur ont évolué au gré des nombreux projets d'analyse distincts les uns des autres. Mais ce que nous constatons, c'est que tout examen exhaustif de l'évolution du cadre de l'analyse des erreurs – de ses débuts timides aux tendances actuelles - conduit inexorablement à un positionnement sur un débat opposant grammaticalité à acceptabilité. Cette opposition a pour but de définir et limiter la portée de l'objet d'analyse - et comme nous l'avons illustré dans les sections précédentes, la position retenue est souvent uniquement du côté de la correction du code linguistique.

Cette approche est tout à fait rationnelle, vis-à-vis des nombreuses définitions avancées de ce qui doit être considéré comme une « erreur ». Si nous pensons par ailleurs à l'apprentissage des langues comme une «activité de classe », cette approche trouve une résonance particulière tant chez les apprenants que chez les didacticiens. Toutefois, si nous envisageons l'apprentissage d'une langue étrangère davantage comme un outil de communication auquel les apprenants peuvent avoir recours « en dehors de la salle de classe » - il est évident que le jugement de grammaticalité ne peut pas être le seul but de la pratique pédagogique. Il ne peut pas non plus en être le seul objectif de l'analyse d'erreur. L'adéquation du genre textuel, la correction pragmatique (cf. section 5.1.3), l'organisation et la structuration informationnelles (à un niveau inter et intra-phrastique), entre autres, doivent également être pris en compte.

En effet, la tendance actuelle montre qu'un nombre considérable de projets d'analyses d'erreurs, effectués en milieu universitaire (voir par exemple, ICLE[19] et TREACLE[20]), se distingue des travaux précédant en adoptant des approches d'analyse qui vont au-delà de l'unité lexicale individuelle voire au-delà de l'unité phrastique individuelle – et donc au-delà de la notion de grammaticalité. Ces approches incorporent dans un premier temps des analyses d'unité lexicale multi-mots ou ce que d'autres appellent des unités de phraséologie lexicale (cf. Osborne 2008 ; Thewissen 2008). D'autres approches s'intéressent à la structure et à la distribution proprement informationnelles : dans ces cas, l'étude de la cohésion et de la cohérence (cf. section 2.3.2) sont de mise et l'analyse porte sur des éléments qui provoquent en principe (i) des problèmes de saillance en termes de rupture dans la continuité référentielle et (ii) des problèmes de continuité sémantique

[19] International Corpus of Learner English (*ICLE*)
[20] Teaching Resource Extraction from an Annotated Corpus of Learner English (TREACLE) est un projet d'analyse de corpus d'apprenants entre deux universités espagnoles : Universidad Autonoma de Madrid et Universidad Politecnica de Valencia.

ou de progression thématique voire (iii) ce que s'apparenteraient à des problèmes de cohérence propositionnelle, selon Carter-Thomas (2009a).

Toutes ces notions différentes permettent d'avoir un aperçu de quelques-unes des considérations textuelles qui apparaissent au fur et à mesure, certes de manière isolée et fragmentée, dans les différentes études portant sur l'analyse des erreurs que nous avons pu consulter. Ce qu'il manque, à notre avis, est un moyen d'examiner l'ensemble de ces points dans un seul projet – de manière à bien mettre en avant la complémentarité de l'aspect d'une part proprement grammatical et d'autre part l'aspect proprement textuel. Et c'est justement ce que nous proposons de faire dans le présent travail. En effet, comme nous le verrons dans la section 4.2.1.4, notre corpus d'apprenants est constitué de textes (i) rédigés par des apprenants d'un niveau dit intermédiaire[21] et (ii) qui doivent s'inscrire dans le genre textuel dit argumentatif ayant des caractéristiques propres aux domaines des sciences économiques (notamment, entre autres, pour ce qui est du registre et le vocabulaire technique). De ce fait, nous pouvons étudier l'acceptabilité de ces textes – aussi bien au niveau micro-textuel qu'au niveau macro-textuel – qui doit répondre à des exigences linguistiques et situationnelles strictes. Notre analyse propose donc d'examiner l'ensemble les erreurs imputables aux systèmes linguistiques (cf. chapitre V) et l'ensemble de celles qui sont imputables à l'acceptabilité textuelle (cf. chapitre VI).

[21] Pour rappel, ce niveau correspond approximativement à B1-B2 dans l'échelle du cadre européen commun de référence pour les langues (CECRL).

(Chapitre II) Genèse d'un courant didactique : focus sur l'unité TEXTE

Après avoir examiné l'aspect épistémologique des erreurs linguistiques, nous passons maintenant aux aspects proprement didactiques. Et bien que certains aspects présentés ne renvoient pas à des objets réellement « enseignés », ils font partie d'une réflexion pédagogique plus large. Cela étant dit, il est tout particulièrement question dans ce chapitre des aspects ne relevant pas des unités lexicales individuelles mais plutôt de l'unité TEXTE comme un ensemble. L'intérêt n'est pas de soutenir que d'un côté il y a des aspects théoriques et de l'autre des aspects didactiques. Il est plutôt question de mettre en perspective ces deux aspects qui sont souvent traités séparément dans la littérature - sans même que l'on fasse le moindre parallèle entre les erreurs commises au niveau de la microstructure du texte (chapitre I) et celles commises au niveau de la macrostructure. C'est justement ce parallèle que nous nous attacherons à présenter dans les sections qui suivent.

Pour faciliter la description de cette macro-unité textuelle, le chapitre est divisé en trois grandes thématiques – sans bien entendu chercher à instaurer une frontière distincte et arbitraire entre chacune d'elles. Dans la première section 2.1, il s'agit tout d'abord de s'intéresser aux effets supposés de la métalinguistique sur la construction du texte, en termes d'avantages qu'apporteraient les connaissances linguistiques explicites. Dans la section 2.2, il est ensuite question de réfléchir aux impacts de la culture, au sens de pratiques rédactionnelles institutionnalisées, sur la rédaction elle-même. Et enfin dans la section 2.3, il s'agit de porter une attention toute particulière aux différents éléments qui donnent au texte sa texture, le rendant de ce fait acceptable en tant que tel.

2.1 Réflexion métalinguistique

Une des principales difficultés qu'engendre l'apprentissage d'une langue étrangère est celle de la production du discours. En effet, bien construire un discours composé de plusieurs énoncés formant un tout complet n'est pas sans obstacles pour l'apprenant-scripteur (ou l'apprenant-écrivain) en langue étrangère. Ceci concerne directement, à titre d'exemple, les étudiants choisissant de se spécialiser ou de poursuivre leurs études supérieures dans une langue et culture discursive autre que la leur et qui témoignent, par la suite, des nombreuses entraves rencontrées à l'écrit – notamment au niveau de la mise en texte, tout au long de leurs différents parcours. Ce problème, qui se révèle inexplicable strictement en termes de connaissances linguistiques générales qu'auraient acquises les apprenants, ne se limite pas aux débutants mais concerne également ceux ayant un niveau dit intermédiaire voire avancé.

Ces constats ne sont pas nouveaux mais continuent à nous faire nous interroger sur les relations effectives qu'entretiennent les connaissances (méta)linguistiques[22] qu'aurait acquises un apprenant avec l'usage réel que ce dernier en fait. En effet, il est généralement admis qu'il y aura forcément une certaine corrélation entre le niveau en langue étrangère d'un apprenant et sa production réelle. Autrement dit, nous nous attendrons à une production de niveau débutant et par voie de conséquence aux erreurs relatives d'un apprenant dit « débutant », et ainsi de suite. Mais qu'en est-il de l'apprenant avancé qui aurait une certaine maîtrise des connaissances (méta)linguistiques en langue étrangère mais qui s'avère incapable de construire une unité textuelle sans *maladresses,* et dont l' « *étrangéité* » se révèle évidente pour tout locuteur natif ? Qu'en est-il de ceux qui connaissent « les règles de grammaire par cœur » mais ne parviennent tout de même que maladroitement à structurer leurs idées et leurs discours face à des textes longs requis dans le cadre universitaire. Cette maladresse qui ne traduit pas formellement une insécurité ou incapacité linguistique nous oblige à réfléchir davantage sur les compétences nécessaires pour une production écrite réussie en langue étrangère.

2.1.1 Connaissances grammaticales explicites

Les deux principales questions qui guident notre réflexion dans cette section sont les suivantes : (i) faut-il être conscient ou capable d'expliciter les règles linguistiques d'une langue pour pouvoir la manier à bon escient (?) et (ii) s'agit-il d'une dynamique d'apprentissage et d'emploi identique

[22] Métalinguistique est entendue ici en tant que connaissances grammaticales explicites ou le fait de pouvoir utiliser la terminologie linguistique pour parler de la langue : que ce soit à un niveau basique ou approfondi. Notons toutefois que le terme métalinguistique a plusieurs acceptations selon le courant et la discipline. Voir Huot, D. & Schmidt, R. (1996).

entre les connaissances linguistiques en langue maternelle et en langue étrangère ? Si la réponse est généralement admise comme étant négative pour la première question, il n'y a pas de consensus général pour la deuxième. En effet, nombreux sont les linguistes et didacticiens (cf. Dabène 1992 ; Gombert 1996 ; Ellis 2006) qui étudient cette problématique dont l'objectif principal est de décrire le rapport entretenu entre connaissances linguistiques explicites et apprentissage d'une langue.

Pour ceux qui s'intéressent à ces questions, la notion d'erreur ou de maladresse – indépendamment du niveau de son occurrence (micro ou macrostructure) – n'est pas posée en tant que telle. L'intérêt central réside alors dans l'étude de toute corrélation possible entre connaissance théorique et pratique réelle ou entre compétence et performance selon la dichotomie chomskyenne. Toutefois la question ne manque pas de susciter un intérêt controversé chez certains (cf. Myhill et al. 2013 ; Thwaite 2015) qui s'intéressent aux connaissances métalinguistiques chez les enseignants et l'incidence directe sur leur pratique pédagogique d'une part, et ceux (cf. Alderson et al. 1997 ; Ellis 2006) qui cherchent d'autre part à établir une corrélation chez les apprenants et leur niveau de production en langue étrangère – ou du moins à en étudier son impact.

Notons que ces nombreuses différences de paradigmes ne présupposent pas une complémentarité entre les différentes approches et donc entre les différents résultats obtenus. C'est effectivement loin d'être le cas. Force est que constater que certaines études, comme par exemple Alderson et al. (1997), mettent en avant le manque de lien entres les connaissances linguistiques explicites et la compétence en termes de production réelle dont la corrélation est dite « faible », tandis qu'elle est dite « modérée à forte » par Roehr (2006). Il convient néanmoins de signaler une contre tendance observée à travers une littérature existante très controversée. En effet, comme nous le montrent Hudson (2001) et Lancaster & Olinger (2014), qui dressent un état des lieux critique des études défavorables à tout métalangage dans l'enseignement des langues, cette tendance dans laquelle l'explicitation des connaissances linguistiques est jugée contre nature et contreproductive s'avère très problématique dans la mesure où les observations varient énormément selon le positionnement théorique, voire idéologique de celui qui se trouve face à cette question épineuse.

Cela dit, nous rejoignons particulièrement Lancaster & Olinger quand ils affirment que la question reste posée et qu'il y a une certaine résistance qui persiste chez les enseignants de langue eux-mêmes qui « sense that their students do benefit from instruction that heightens their awareness of the ways the details of language works in texts » (2014 : 1). Cela étant dit, nous avons voulu montrer, par le biais de cet exposé succinct, à quel point une question a priori anodine peut s'avérer riche en réflexion. Toutefois, nous soulignons notre attachement tout d'abord en tant qu'enseignant

à une pédagogie dans laquelle les connaissances linguistiques ont toute leur place – et ce, qu'il s'agisse des cours de LANSAD[23] ou de langue de spécialité[24], notamment face à un public qui d'ordinaire en fait la demande. Et dans un deuxième temps nous soulignons en tant que linguiste, la nature subjective entre les différentes méthodes utilisées pour mesurer l'impact de l'ensemble des connaissances théoriques acquises et l'utilisation réelle qu'un apprenant en fait. Notons également que nous sommes conscient qu'il est possible d'apprendre une langue autre que sa langue maternelle sans métalangage aucun et qu'il est tout aussi possible que le métalangage fournisse à d'autres un moyen d' « accéder » à l'utilisation d'une langue étrangère. Face à ces observations, l'incidence directe des connaissances métalinguistiques sur l'objet texte reste à définir.

2.1.2 Maturité syntaxique

Contrairement à ceux qui s'intéressent à l'impact des connaissances linguistiques explicites sur la production écrite, d'autres cherchent un moyen objectif d'évaluer l'objet texte lui-même – et d'y identifier des observables concrets permettant d'évaluer le niveau réel des apprenants-scripteurs : autrement dit, certains utilisent le texte lui-même comme objet et mesure d'étude. C'est notamment le cas de Hudson (2009) et Arnaud (1984) qui étudient non seulement l'application des règles syntaxiques dans un texte donné mais tout particulièrement la densité voire la complexité à la fois lexicale et syntaxique des phrases employées. Le raisonnement derrière cet angle d'approche a été succinctement résumé par Péry-Woodley :

> Si la maturité syntaxique va de pair avec l'apprentissage de l'écrit, et par conséquent avec le progrès et le bien écrire, ce n'est pas parce que les phrases complexes sont une marque de qualité, mais parce que les stratégies d'écriture de haut niveau nécessitent une syntaxe complexe. (1993 : 29)

Cela étant, un des tournants majeurs dans l'évaluation d'un texte comme unité à part entière se trouve dans une mesure mise en avant par Kellogg Hunt : une mesure qui fut progressivement adoptée dans de nombreux projets d'évaluation de textes d'apprenants. Ce dernier a identifié ce que l'on appelle désormais des « minimal terminable units » ou « T-unit » et il en a fourni la définition suivante « one main clause plus whatever subordinate clauses happen to be attached to or embedded within it » (Hunt, 1965 : 305). Cette définition a été davantage explicitée par Degand & Hadermann (2009 : 21-22) qui soulignent que la « T-unit divise les phrases indépendantes coordonnées […] mais regroupe la principale de ses subordonnées.

[23] Langues pour Spécialistes d'Autres Disciplines, entendu comme des cours de langues générales.
[24] Entendu comme des cours propre à un domaine, par exemple un cours d'anglais juridique.

Ainsi pour Hunt (1965) la T-unit fournit une mesure fiable face à l'ancienne tradition de prise en compte à la fois du nombre et de la longueur des propositions et des phrases en guise d'indice de complexité voire de maturité d'un texte donné. Ces pratiques jugées trop arbitraires ont laissé place à un outil qui dorénavant selon Degand & Hadermann (2009) « permet de déterminer plus objectivement la complexité discursive en uniformisant la mesure de la longueur et de la complexité syntaxique ».

En effet, pour ce qui est de son application en langue étrangère, Polio (1997) repasse en revue des études dans lesquelles la T-unit a été adaptée aux analyses portant spécifiquement sur les erreurs d'apprenants en langue étrangère. Cette dernière souligne qu'en plus de la décomposition nécessaire en « minimal terminable unit », des chercheurs ont tenté d'intégrer le nombre de T-unit dépourvu d'erreurs dans leur calcul de la maturité syntaxique des apprenants – obtenant des résultats quantitatifs très probants. Toutefois elle souligne que les différents résultats obtenus sont tous à relativiser en raison (i) des différences d'appréciation dans la notion d'erreur et (ii) tout singulièrement, du manque ou de l'absence de tests d'accord inter-annotateurs (cf. section 4.1.2.3 pour une présentation détaillée). Ce qui a pour effet de ne pas savoir dans quelle mesure les résultats sont reproductibles ou généralisables sur une population d'étude similaire.

Nonobstant, notons qu'en dépit des nombreuses critiques sur la T-unit (cf. Lutkus 1987) la mesure de Hunt a permis de remettre en cause la tradition dominante – qui existait jusqu'alors – d'évaluation de la maturité syntaxique d'un texte aussi bien en langue maternelle qu'en langue étrangère. De plus, tout en reconnaissant certaines limites de la mesure, elle demeure toujours appréciée par ceux qui s'intéressent au développement syntaxique des apprenants et qui souhaitent obtenir des données avant tout quantitatives. De nos jours, la mesure est encore employée et se trouve souvent croisée avec d'autres méthodes, à l'exemple de la progression thématique, comme en témoigne l'étude de Hernandez et al. (2013).

2.2 Réflexion sur les pratiques discursives culturelles

Il convient maintenant de signaler un courant qui s'est détourné des notions de grammaticalité ou de maîtrise lexicale – en termes de connaissances grammaticales, la densité lexicale ou la maturité syntaxique – pour s'intéresser à la pratique rédactionnelle des apprenants elle-même. Ce courant de pensée qui a adopté un ancrage avant tout sociolinguistique est présenté ci-après.

2.2.1 Rhétorique contrastive

Une autre perspective de l'époque, à savoir à partir des années 1960, consistait à concevoir les erreurs d'apprenants en langue étrangère, non pas dans leur configuration strictement phrastique, mais de manière macro-textuelle. Autrement dit ramener l'erreur à un dysfonctionnement textuel, en termes d'écarts par rapport à une norme discursive attendue. Ici, les connaissances linguistiques de l'apprenant ne suffisent plus pour expliquer les erreurs dans le texte écrit en langue étrangère. En effet, Robert B. Kaplan (1966) a été un des premiers à souligner ce problème en soutenant que dans de tels cas nous n'avons plus affaire à un problème de langage en soi, mais de *rhétorique*[25], qui selon ce dernier n'est pas un principe universel.

> Logic […] which is the basis of rhetoric, is evolved out of a culture; it is not universal. Rhetoric, then, is not universal either, but varies from culture to culture and even from time to time within a given culture. It is affected by canons of taste within a given culture at a given time. [ibid : 12]

Il s'ensuit que l'hypothèse conductrice de Kaplan (1966, 1971) a été de soutenir qu'il y avait une différence non-négligeable particulièrement identifiable au niveau de l'organisation textuelle. Et ce, qu'il s'agisse d'une production dite linéaire (directe comme il le prétend pour l'anglais), en forme de parallélismes (pour les langues sémitiques), circulaire (donc indirecte pour les langues orientales) ou encore digressive (pour les langues latines, notamment l'espagnol et le français) relevant de l'imaginaire culturel dans lequel les scripteurs puisent l'organisation informationnelle, tel un moule à discours.

A posteriori, bien que ces catégorisations aient été fortement critiquées sans pour autant être formellement démenties, l'hypothèse sous-jacente qui a été « la pierre angulaire » des théories de *rhétoriques contrastives* reste intacte, à savoir l'existence dans un sens large de ce que l'on pourrait appeler *un moule discursif* qui varie d'une langue à une autre. De plus, la structure rhétorique joue ici un rôle considérable au même titre que la cohésion et la cohérence (*cf.* section 2.3.2) dans la recevabilité du produit texte. En effet, il est à préciser qu'une condition nécessaire à la bonne réalisation textuelle serait la capacité de l'apprenant-scripteur à intégrer l'ensemble des paramètres discursifs culturels lors de la production de nouveaux textes.

Intervient alors la question de savoir comment étudier de façon objective le texte final, dans sa conception macrostructurale, tel qu'il est produit par l'apprenant étranger dans la mesure où celui-

[25] Entendu ici comme englobant les conventions socio-discursives permettant de rédiger des textes différents selon les modalités stylistiques, informationnelles et organisationnelles attendues dans une communauté discursive donnée par opposition à l'*art rhétorique* communément admis comme l'art d'argumenter.

ci heurterait la sensibilité des enseignants-correcteurs qui n'y retrouveraient pas *« les structures rhétoriques attendues »* et qui poseraient ainsi un problème de *« compréhension »* et *« d'appréciation »* voire éventuellement d'évaluation. Il est à noter que dans de tels cas, comme l'a souligné Kaplan, ni le correcteur ni le lecteur potentiel ne reconnait les conventions discursives dans lesquelles l'autre s'inscrit, et ceci a pour conséquence de provoquer des problèmes qui ne sont généralement pas élucidés à point opportun.

Ce problème de la non-utilisation des paramètres discursifs culturels se produit malencontreusement au détriment des rédactions de l'apprenant-scripteur étranger qui ne saurait pas corriger ses écarts discursifs rédactionnels sans prendre en compte les nouvelles conventions scripturales et sans s'y adapter. De ce fait, ce dernier pourrait continuer à écrire un texte considéré *inacceptable* pour la culture discursive dans laquelle il se trouve. En effet, il s'ensuit que l'apprenant risque de se heurter à un problème de fossilisation textuelle liée à la non-reconnaissance des nouveaux modèles à produire, ce qui pourrait l'empêcher de prendre conscience de la nouvelle structure rhétorique en usage et par extension de l'acquérir. D'après notre propre pratique pédagogique, il convient de souligner ici que ceci est souvent un problème aussi bien pour les débutants que pour les apprenants avancés.

2.2.2 L'ancrage des styles intellectuels institutionnalisés

La même question de l'acceptabilité d'un texte est expliquée au moyen d'autres termes par Kramsch (1991) qui rejoint l'argument principal de Kaplan (1966) et celui de Moirand (1979), en identifiant dans le discours « le mode de raisonnement » inculqué par l'institution sociale au sens large :

> Intellectual styles or patterns of thought are socially and culturally determined and are so inseparable from the informational content that communication breakdowns occur more often than not at the level of discourse, not at the level of the facts presented. (1991 : 226)

Pour Kramsch le résultat final reflète les politiques des institutions sociales qui influencent, de manière considérable, la façon donc les langues étrangères sont enseignées et par voie de conséquence les résultats qui en sont tirés. Autrement dit, l'enseignement-apprentissage pourrait, dans une grande mesure, refléter les intérêts des politiques linguistiques du pays concerné : par exemple dans la logique de la tradition nord-américaine où l'accent est mis sur les aspects *culturo-pragmatiques* et *l'usage approprié*, au détriment d'une réflexion intellectuelle et littéraire comme le voulait la tradition européenne de l'époque (*op.cit.* 1991)

Toutefois, Moirand maintient que produire en langue étrangère supposerait dans un premier temps la reconnaissance des modèles que l'on cherche à reproduire : *« on ne peut produire […] les types d'écrit concerné […] avant d'en avoir « vu » dans la langue que l'on apprend»* (1979 : 96). Ce qui signifie qu'il y aurait des modèles distincts d'une langue à une autre. Une hypothèse qui s'inscrit dans la lignée non revendiquée de la rhétorique contrastive.

Il convient néanmoins de comprendre que comparer et contraster des langues se révèle résolument insuffisant dans la mesure où l'on s'intéresse à l'acquisition et l'output final des apprenants en langue étrangère. Ainsi faudra-t-il reconnaître qu'écrire en langue maternelle de même qu'en langue étrangère ne relève pas d'une compétence linguistique que l'on pourrait désigner strictement comme étant du type « lexico-syntaxico-pragmatique[26]» ni d'une compétence discursive strictement culturelle. De ce fait, il nous parait judicieux de privilégier dans la présente étude une approche où l'intérêt serait de localiser, identifier et ensuite analyser les erreurs, non principalement par leur fonction linguistique ou macrostructurale, mais par rapport à l'environnement immédiat dans lequel elles puisent leurs fonctions. Fonction qu'il convient encore de définir.

2.2.2.1 L'apport de Clyne & de Mauranen

Dans la lignée non revendiquée des théories de rhétorique contrastive de Kaplan (1966), on trouve deux chercheurs dont les travaux mettent en avant le caractère singulier de l'institutionnalisation de la culture discursive. En effet, bien qu'ils travaillent depuis plus de vingt ans avec des approches et des langues différentes du point de vue de leur racine linguistique, Clyne (1981, 1987, 2002) et Mauraunen (1993a, 1993b, 1996) ont démontré à quel point les paramètres culturels sont importants dans l'appréciation d'une production textuelle comme un ensemble construit. De plus, en dépit du fait qu'ils ne s'intéressent pas particulièrement à la notion d'erreur ou de maladresse dans des productions écrites, l'ensemble de leurs travaux ont des répercussions tant en didactique des langues maternelles qu'en didactique des langues étrangères.

Commençons d'abord avec Michael Clyne. Pour ce dernier, qui s'est intéressé en 1981 au rapport établi entre la structure discursive et la culture – majoritairement en allemand et anglais[27] langues maternelles, il y a des différences non-négligeables au niveau de (i) la linéarité en termes d'organisation du discours, (ii) de la prolixité, (iii) du formalisme en termes de règles d'adhérence

[26] Cf. 2.3.3 pour une discussion sur les compétences nécessaires pour une rédaction réussie en langue étrangère.
[27] Il s'agit de l'anglais britannique et australien que l'auteur différencie de l'anglais américain – de par l'approche dans l'enseignement aux Etats Unis qu'il juge différente et les attentes évaluatives. (cf. Clyne 1981, 1987).

strictes à des conventions de rédaction et enfin (iv) du rythme – au sens de tour de paroles. Pour l'auteur, l'écrit scolaire ou institutionnalisé relève donc du « culturally-conditioned formalism », ce qui explique en partie selon ce dernier que certains étudiants étrangers échouent dans des établissements scolaires dans des pays hôtes avec une langue d'apprentissage différente de la leur – tout particulièrement parce que l'ensemble des règles ou des formalismes culturels propre au discours ne leur ont pas été explicités. Les apprenants peuvent également rencontrer par la suite des difficultés à intégrer la communauté discursive – au sens large – de leur pays d'accueil.

Ce problème ne se limite pas à des étudiants et peut donc affecter des « écrivains-scripteurs experts ». Clyne (1981) rapporte par exemple qu'une traduction anglaise de l'œuvre « Soziolinguistik » de l'allemand Norbert Dittmar a été vivement critiquée en anglais pour son aspect « chaotique », son absence « de cohésion » et ses nombreuses digressions. Mais à en croire Clyne, tout cela n'est que le fruit d'une méconnaissance des structures discursives en vigueur en lange allemande. Suite à ces premiers résultats, l'auteur a multiplié les études contrastives anglais-allemand dans le but d'identifier davantage, de façon concrète, les différents éléments qui pourraient heurter le passage d'une communauté discursive à une autre.

En 1987, l'auteur s'est de nouveau intéressé à ce phénomène en approfondissant davantage les catégories issues de ses précédents travaux. Des écarts sont alors observés dans (i) les normes rédactionnelles (aussi bien en termes de produit final que de regards évaluatifs correspondants), (ii) le formalisme du registre, (iii) les patrons discursifs, jusqu'à la récurrence de certaines (iv) structures grammaticales et (v) rhétoriques. Ces éléments de divergences entre les deux cultures discursives sont passés au peigne fin les uns après les autres. Les différences sont telles que Clyne (1987) souligne que « mastery of discourse conventions » doit être un pré-requis pour pouvoir naviguer à travers « the international academic scene ». Ce faisant, il fait écho à son alerte lancé en 1981 :

> If culture-specific discourse structures really play an important role, they should occupy a prominent place in teaching programs for second and foreign languages, including languages for special purposes. (1981 : 63)

A travers ses différentes publications, l'auteur s'est attaqué aux différences discursives dans de nombreuses langues – souvent en les comparant à l'allemand et l'anglais – jusqu'en 2002, quand il a commencé à militer pour davantage de ce qu'il appelle « contrastive discourse studies ». L'objectif selon lui n'est pas de transformer à tout prix les traits culturels distincts en objet enseignables mais plutôt de permettre une meilleure sensibilisation et par la suite, une meilleure disposition à la communication interculturelle. Il met en garde cependant contre le fait de vouloir

enseigner aux apprenants à écrire avec des valeurs discursives qui ne sont pas les leurs. Par exemple, – avec le principe rapporté par une étude de Mauranen (1993a) – Clyne (2002) souligne le fait qu'il y a des cultures où l'accent est mis sur la politesse inversée (par exemple en finnois), où il est impoli de constamment rappeler les étapes à venir dans un travail écrit.

Il convient alors de souligner que les différences observées dans les travaux de Clyne n'ont pas pour but d'indiquer un moyen de corriger les erreurs ou maladresses interculturelles mais visent plutôt la sensibilisation culturelle : nous pensons que cela aurait pour effet de permettre aussi bien à l'enseignant qu'à l'apprenant de mieux comprendre et ainsi orienter leur discours respectifs selon le type de public visé. Ainsi, à titre d'exemple, les nombreuses digressions observées chez les uns, les structures asymétriques de paragraphe ou la linéarité discursive chez les autres ne seront pas forcément compris comme étant des erreurs macrostructurales selon la culture discursive de l'apprenant-scripteur. Mais la prise en compte de la différence discursive de l'apprenant et la culture discursive dans laquelle celui-ci souhaite s'intégrer devrait alors être davantage prise en compte dans l'enseignement et l'évaluation des productions écrites.

Pour ce qui est des travaux d'Anna Mauranen (1993a, 1993b, 1996), l'orientation est sensiblement similaire à celle de Clyne. Mais à la différence de ce dernier, Mauranen s'intéresse tout particulièrement à l'enseignement et l'utilisation effective de l'anglais en tant que langue étrangère et notamment en tant que lingua franca dans la communauté scientifique au sens large. De plus, une grande majorité de ses travaux s'articulent autour des comparaisons entre l'usage de l'anglais dans le monde dit académique chez des anglophones natifs et chez la communauté universitaire finnoise dont elle est issue. On pourrait alors ramener la problématique générale d'une grande partie de ses recherches à la citation suivante.

> Why it is important to show that cultural differences permeate the world of science as well as any other aspect of social and intellectual behaviour is because, as we all know, cultural differences are most problematic when the users are not aware of them. Practically any native speaker of a particular language can tell a foreign writer from errors or peculiarities in lexis and grammar. However, even if such mistakes are eliminated from a text, a number of "foreign" features are left at the discourse level, which affect our comprehension and assessment of a text, although we are usually not aware of them. (1993a :157-8)

Il s'ensuit alors pour Mauranen que des caractéristiques proprement rhétoriques et ou discursives sont profondément ancrées dans la valeur discursive de la culture d'appartenance de celui qui s'exprime – ceci est notamment le cas pour les éléments qui se trouvent « above the level of the

sentence » (1993a : 157). Cependant, au lieu de militer pour une sensibilisation des différents paramètres culturels identifiés dans le discours, elle met en garde contre le risque de vouloir standardiser le discours académique avec un modèle issu d'une culture discursive nationale : ce qui aurait pour conséquence de freiner la réception des travaux de ceux qui n'adhèrent pas au modèle dominant. Ce faisant, l'auteur pense que l'on contribuera à la préservation de « small cultures for no other reason than to keep up diversity » et donc des « smallish, local academic communities with their own discourse and rhetorical practices » (ibid. : 172).

Cette approche de Mauranen[28] – qui renvoie d'une certaine manière à la notion de langue dominante par opposition à langue dominée ou encore à l'idée que le fait d'imposer un modèle unique de rédaction traduise une sorte d'uniformisation du modèle d'expression et de la pensée – laisse inévitablement entendre la mort programmée des communautés discursives qui se trouvent en position minoritaires à travers le monde. Ce constat préoccupant, qui est partagé – entre autres – par Judet de la Combe et Wismann (2004), a toutefois le mérite de montrer l'existence de véritables différences culturelles identifiables dans les productions écrites – puisqu'on arrive même à provoquer une réaction de crainte de voir les pratiques discursives nationales affectées par l'hégémonie d'un modèle venant d'ailleurs et dont les valeurs rhétoriques ne seraient pas partagées (cf. Kuteeva & Mauranen 2014).

Pour ramener alors le débat des pratiques institutionnalisées à notre contexte d'étude, il devient alors évident que la notion d'erreur au niveau de l'objet texte – qu'il soit en langue maternelle ou en langue étrangère – peut s'avérer différente selon le contexte à la fois textuel et culturel de la rédaction, sans oublier bien entendu le public visé avec les valeurs d'évaluations qui peuvent être le propre d'une communauté discursive. Il convient par conséquent de souligner que l'on doit prendre l'ensemble de ces paramètres en compte dans l'enseignement et l'évaluation des langues étrangères. Plus précisément, on s'attendra par exemple à ce que le public francophone dans des cours d'anglais général ou de LANSAD soit sensibilisé dans une moindre mesure aux différences macro-textuelles que l'on peut trouver dans des textes rédigés en anglais par des anglophones natifs, ne serait-ce que de l'ordre organisationnel et argumentatif. Tandis que ceux qui s'inscrivent

[28] Notons, à titre d'information, que les travaux récents de Mauranen (2006, 2012) s'intéressent de plus en plus à l'utilisation de la langue anglaise non seulement en comparant les discours produits par des non-natifs à ceux des locuteurs dits natifs, mais tout particulièrement en comparant les discours des non-natifs entre eux. De même, elle continue à étudier le rapport de force de l'anglais en tant que *lingua franca* (ALF) : à la fois (i) par rapport à son emploi réel chez des universitaires qui doivent l'utiliser pour atteindre un public large et (ii) par rapport à son développement qui – à son tour – commence vraisemblablement « à façonner » l'usage de l'anglais académique général. Notons également qu'elle dirige aujourd'hui le premier projet international qui vise la constitution d'un corpus d'ALF : à savoir « *English as a Lingua Franca in Academic Settings (ELFA)* ».

en licence d'anglais ou des cours de langue de spécialité (par exemple, *English for Aeronautical Engineering*) – dont l'un des objectifs est de les amener à s'intégrer dans une nouvelle communauté discursive – peuvent s'attendre véritablement à ce que la sensibilisation dépasse le simple stage informationnel et qu'il y ait une sorte de mise en pratique. Et ce, afin d'éviter toute forme d'erreur ou maladresse observable au niveau proprement macro-textuel.

2.2.2.2 L'écrit comme *reader-oriented* ou *writer-oriented*

Hormis le fait de souligner les différences culturelles dans l'organisation macro-structurelle d'un texte écrit, certains vont jusqu'à affirmer l'existence d'un caractère proprement institutionnalisé identifiable au niveau de la présentation argumentative : ou plus précisément dans l'avancement des contenus informationnels et de la façon dont ces éléments sont reliés entre eux. Pour ceux qui adoptent cet angle d'analyse, le focus est soit sur le lecteur potentiel dans la mesure où la responsabilité du scripteur est d'aider ce dernier à suivre le bien-fondé de son raisonnement, soit sur l'écrivain-scripteur lui-même où l'objectif est d'apprécier l'ensemble de ses connaissances élargies sur un point précis. Ces deux perspectives se trouvent notamment dans les travaux de Siepmann (2006). Cependant, sans entrer dans un débat théorique – en comparant les différentes approches et perspectives mises en avant chez ceux qui travaillent sur ces points précis[29]–, nous limiterons notre discussion ici à la simple présentation des principaux tenants.

En effet, certains chercheurs (cf. Siepmann 2006 ; Clyne 1987 ; Galtung 1981 ; Kaplan 1966) mettent en avant le fait qu'un texte puisse être évalué différemment selon la culture discursive : soit (i) en valorisant la précision de l'information présentée dans un sujet traité de manière circonscrite, sans oublier bien entendu les enchaînements dans l'argumentation en rapport direct avec le sujet et la force de conviction ou de persuasion de l'auteur lui-même, soit (ii) en n'accordant que très peu de valeur à la forme ou au style de la rédaction et en ne jugeant le texte que par l'abondance des connaissances directes et périphériques apportées au sujet traité. Il va sans dire que nous jugeons ces deux points complémentaires, mais force est que constater que les auteurs cités ci-dessus, entre autres, soutiennent que certaines cultures discursives se trouvent aux deux extrémités de cet axe. Comme par exemple l'anglais pour le premier cas de figure et l'allemand pour le deuxième, comme nous le rapportent Siepmann et Clyne.

[29] Ce sujet est traité majoritairement en sciences de l'éducation et des études interdisciplinaires, mais ne constitue en soi un objet d'étude à part entière en linguistique – du moins dans la littérature que nous avons pu consulter.

Bien que ces arguments nous paraissent trop caricaturaux en l'état, nous notons que le principe de la rédaction dite « *reader-oriented* » est très largement repris[30] dans l'enseignement de l'anglais langue étrangère dans de nombreuses universités non anglophones à travers le monde : et le raisonnement derrière cette approche se trouve justement dans la perspective contrastive que la communauté scientifique anglophone (traitée dans ces littératures comme un ensemble homogène) privilégierait la linéarité du texte, la clarté de l'écrit et la prise en compte du lecteur dans une sorte de dialogue entre le scripteur et le lecteur. A l'opposé donc de ce continuum, les écrits dits « *writer-oriented* » en anglais se trouveront alors critiqués comme étant des textes mal-écrits voire un peu « chaotique » pour reprendre l'exemple de Clyne (1981) (cf. section 2.2.2.1).

Cela étant dit, en présentant cet aspect supposément culturel dans l'évaluation des productions écrites – l'objectif est tout simplement de souligner à quel point la notion d'erreur ou de maladresse repérée au niveau proprement macro-textuel peut s'avérer délicate, pour ainsi dire. De ce fait, on pourrait alors soutenir d'une part qu'au niveau du lexique ou de la grammaire (en termes de syntaxe) les règles sont plus au moins connues – et qu'il y a un certain consensus sur ce qui constitue une erreur tandis qu'on est encore loin d'avoir un consensus sur ce qu'est une erreur macro-textuelle.

2.3 Réflexion sur la textualisation

Etant donné que nous avons passé en revue la notion d'erreur en termes de connaissances (méta)linguistiques (cf. section 2.1), en termes d'écarts par rapport aux pratiques discursives institutionnalisées (cf. section 2.2), passons maintenant aux différents types d'erreurs que l'on peut signaler, à proprement parler, au niveau de la construction du texte. Tout d'abord, si nous admettons que le texte se réalise, dans sa forme la plus appauvrie, par une suite de phrases, il devient alors parfaitement compréhensible que notre première constatation porte sur le caractère composé du texte. En effet, ce caractère composé suppose entre autres l'existence de règles et de structures sous-jacentes, relatives à la bonne réalisation textuelle – le tout assujetti bien entendu à l'environnement institutionnalisé de la rédaction. Nous appellerons ce caractère composé la « mise en texte » ou la « textualisation ».

Mais encore faut-il pouvoir définir concrètement ce processus en termes de structures et principes combinatoires, ce qui supposerait d'emblée d'entrer dans une problématique épistémologique où

[30] Cf. par exemple Busch-Lauer (2002) où elle préconise de passer à l'approche dite « *reader-oriented* ».

tout chemine vers les entours relatifs du texte en tant que produit composé – avec son propre processus de construction et ses éléments constitutifs. Il ne serait donc pas dérisoire de rappeler ici que notre objet d'étude peine à recevoir et à admettre une définition qui conviendrait à tous, notamment en raison du fait que l'évaluation de l'objet final peut être assujettie à des critères culturels – sans oublier les disparités terminologiques[31] existant chez les uns et les autres qui font du texte (ou de la production écrite aussi bien en langue maternelle qu'en langue étrangère) un objet de prédilection. Cela étant dit, intéressons-nous brièvement maintenant à l'impact de l'environnement institutionnalisé sur la production écrite et ensuite portons une attention particulière aux éléments constitutifs qui pourraient faire défaut à nos apprenants-scripteurs lors de leur mise en texte.

2.3.1 L'influence des genres textuels

Si nous reconnaissons qu'il existe des communautés discursives, il conviendrait également d'admettre deux choses supplémentaires: (i) d'une part que ces dernières sont multiples et (ii) d'autre part qu'il y a des caractéristiques distinctes nous permettant de les différencier. Mais tout cela demeure dans une certaine mesure très théorique, notamment en raison du fait que « communauté discursive » n'est pas interchangeable avec « communauté linguistique ». Ces différents constats qui appellent à réfléchir constituent, en quelque sorte, le point de départ de nombreuses études (cf. Clyne 1987 ; Swales 1990 ; Malrieu 2004). Et parmi ceux qui travaillent sur la notion de pratiques et communautés discursives, beaucoup accordent une attention spécifique au phénomène qui nous intéresse ici, à savoir, l'influence de ce que certains appellent le genre ou des genres textuels sur la production écrite. Toutefois dans un souci de brièveté, nous allons simplement souligner pourquoi ce point est important lors de l'évaluation des productions écrites, sans entrer dans la présentation des débats terminologiques sur ce qu'est un genre.

Cela dit, signalons tout de même deux citations de John Swales dont les travaux constituent un socle dans l'analyse des genres dans la littérature anglo-saxonne[32] – notamment chez ceux qui travaillent sur les langues de spécialité. Nous pensons que ces deux définitions permettront de dessiner le contour général de la conception du genre chez Swales et de ce fait la conception que nous adoptons dans la présente étude.

---

[31] En complément de la section 2.2, précisons que la notion de texte lui-même diffère selon l'approche employée. Deux courants majeurs s'opposent, sans en être complémentaires. Voir par exemple Adam (1993, 2003, 2004) et Halliday & Hasan (1976) ou encore Peytard & Moirand (1992) dans une moindre mesure.

[32] Le choix de littérature est purement pragmatique. L'analyse de genres dans la littérature francophone rassemble des acceptions du terme « genre » venant de la tradition des études littéraires françaises et ne constitue pas en soi des cadres d'analyses complémentaires ou transposables à la langue anglaise. (cf. la définition de Malrieu 2004).

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community. (1990 : 58)

La deuxième définition est davantage circonscrite.

> As I see it, the work of genre is to mediate between social situations and the texts that respond strategically to the exigencies of those situations. (2009 : 14)

Ce qu'il faut retenir de ces deux citations est qu'à l'intérieur des communautés discursives, il y a des genres différents. De plus, la structure schématique, le contenu, le style et les situations sociales (que l'on pourrait traduire par contexte de rédaction) renvoient à un certain nombre de contraintes imposées au genre : pour adhérer donc à une communauté discursive et un genre textuel spécifique le scripteur n'a d'autres choix que de respecter ces contraintes. Il en va de même si l'on considère les productions écrites en langue étrangère - tout particulièrement celles de notre corpus d'étude - comme appartenant à un genre que l'on pourrait appeler universitaire, le non-respect de ces contraintes constituerait un motif valable de rejet de l'ensemble textuel ou du moins d'évaluation peu favorable. Le genre a donc un rôle important à jouer dans l'appréciation de l'objet texte comme un ensemble complet.

2.3.2 Cohérence et cohésion

Examinons maintenant un autre composant indispensable au processus de textualisation. Comme nous l'avons souligné, dès lors que l'on s'intéresse à l'ensemble textuel, l'objectif même de l'analyse linguistique traditionnelle change. On s'éloigne du syntaxique et du lexical, à proprement parler, pour aller vers « le discursif » ou ce qui fait l'unité de l'objet TEXTE. Unité qui se définit de maintes façons selon le courant linguistique d'ancrage. Sanders & Pander Maat nous rapportent que :

> Discourse is more than a random set of occurrences: it shows connectedness. A central objective of linguists working on discourse level is to characterize this connectedness. (2006 : 591)

En effet, c'est ce « connectedness[33] », traduit par *connexité*, qui intéresse surtout les spécialistes travaillant sur l'objet texte (dans un courant appelé la linguistique textuelle) en raison du fait qu'il permet de différencier entre autres des phrases individuelles et des phrases qui réalisent un ensemble textuel plus large. On comprend donc pourquoi cette idée s'est répercutée dans les études des textes écrits aussi bien en langue maternelle qu'en langue étrangère. L'étude de cette notion de connexité « a vraisemblablement commencé à gagner du terrain » après la publication de l'ouvrage *Cohesion in English* par Halliday & Hasan en 1967. L'ouvrage qui avait pour but de décrire cet épiphénomène de l'époque est devenu par la suite un objet de référence pendant plus d'une quarantaine d'années. Nous notons toutefois, malgré la prévalence de cet ouvrage et de ses lignes directrices, que le terme *cohésion* n'englobe pas de manière satisfaisante le principe de « connectedness » et continue d'évoquer des paradigmes variés chez les différents linguistes. En bref, dans la tradition de Halliday & Hasan, son acception privilégie des « non-structural relations above the sentence » : des relations qui sont organisées en cinq catégories. Le tableau ci-après illustre succinctement l'étendue de cette taxonomie[34].

Cohésion grammaticale	• Référence • Ellipse • substitution • conjonction
Cohésion lexicale	• Cohésion lexicale I. Répétition d'items lexicaux II. Synonymie ou quasi-synonymie /y compris l'hyponymie III. Collocation

Tableau 4 : La cohésion selon Halliday & Hasan (adaptée par Martin 2003)

Les études qui se sont inscrites dans cette lignée se sont bien entendu focalisées sur ce phénomène dans la logique conceptuelle de Halliday & Hasan. Il en ressort cependant que la cohésion, telle que définie ci-dessus, n'est pas la seule condition nécessaire pour la connexité globale d'un texte selon Sander & Pander Maat (2006). Autrement dit, elle ne constitue pas en elle-même la totalité des moyens qui doivent être mis en place pour réaliser les connections et liages nécessaires pour qu'un écrit soit perçu comme étant un ensemble textuel uni. A titre d'illustration, la cohésion ne permet donc pas d'élucider de manière convaincante l'ensemble des dysfonctionnements « non structural » dans le texte, selon la terminologie hallidayienne. Nous jugeons alors inévitable de conclure que cette approche est incomplète et mérite une révision.

[33] Cf. Charolles 1986 ; Adam 1993

[34] Nous jugeons judicieux de pas nous attarder sur la définition de ces sous-catégories en raison du fait que cela ne relève pas de notre propos et également parce qu'on ne les exploitera pas en détail dans la présente étude. Pour une présentation détaillée, nous renvoyons à Halliday & Hasan (1967) ; ou Martin (2003)

Il est toutefois à noter que certains (cf. notamment Charolles 1978 ; Adam 1993 ; Maingueneau 1998) ont très rapidement écarté la définition de Halliday, au sujet de la « non-structural relations above the sentence » : en raison notamment du fait que l'analyse déduite en langue anglaise, des textes anglais, ne corroborent guère celle faite dans d'autres langues, notamment le français pour ce qui est de notre ressort. Certains ont donc proposé une définition distincte, comme Maingueneau dans l'exemple suivant.

> Ces phénomènes de reprises contribuent de manière essentielle à lier entre elles les phrases d'un texte, à lui donner sa cohésion tout en le faisant progresser. (1998 : 171)

Ce dernier propose d'expliquer la cohésion uniquement par les rapports de reprises (ou de coréférence) entretenus dans le texte : les reprises privilégiées sont d'ordre anaphorique, cataphorique et endophorique. D'un autre coté, d'autres, comme Charolles (1978), jugeant la notion de cohésion insatisfaisante pour réaliser la connexité d'un texte, ont avancé des arguments pour l'inclusion de la cohésion dans ce qu'ils appellent plus largement la cohérence.

> […] il ne semble plus possible techniquement d'opérer une partition rigoureuse entre les règles de portée textuelle et les règles de portée discursive. Les grammaires de texte font éclater les frontières généralement admises entre la sémantique et la pragmatique, entre l'immanent et le situationnel, d'où, à notre avis, l'inutilité présente d'une distinction cohésion-cohérence que d'aucuns proposent en se fondant justement sur un partage précis de ces deux territoires. (1978 : 14)

Les arguments de Charolles (1978) ont ouvert la voie à un nouveau cadre d'analyse du produit TEXTE, dans lequel il serait désormais possible de « *dériver des jugements théoriques dits de cohérence recouvrant […] le champ des appréciations vernaculaires à disqualification maximale et des 'jugements de non-standardisation' correspondant aux dépréciations de surface* » (1978 : 8). En effet, Charolles estime que l'étude de la cohérence permet de rendre compte d'un *certain nombre de conditions, tant linguistiques que pragmatiques, qu'un texte doit satisfaire pour être admis comme bien formé (par un récepteur donné, dans une situation donnée).*Selon ce dernier, ces conditions sont les quatre « règles de bonne formation textuelle » et il préconise que son cadre d'analyse permet de relever les dysfonctionnements à l'échelle globale du texte. Ces quatre règles sont présentées ci-après (ibid. : 14-32) :

- MR.I : pour qu'un texte soit cohérent, il faut qu'il comporte dans son développement linéaire des éléments à récurrence stricte.
- MR.II : il faut que son développement s'accompagne d'un apport sémantique constamment renouvelé.

- MR.III : il faut que le développement n'introduise aucun élément sémantique contredisant un contenu posé ou présupposé par une occurrence antérieure ou déductive de celle-ci par inférence.
- MR.IV : dans une séquence ou un texte, il faut que les faits qu'ils dénotent dans un monde représenté soient reliés.

Malgré l'avancée indéniable de ce cadre, il reste perfectible. Les deux premiers éléments qui constituent ce recueil, à savoir définis comme étant *les méta-règles (MR) de répétition et de progression,* peuvent facilement prêter à confusion dans la mesure où ils renvoient tous deux à ce que l'on appelle communément l'analyse de la cohésion. De plus, les deux derniers – *les méta-règles de non-contradiction et de relation* – permettent une marge importante d'interprétation subjective de la part de l'analyste. Toutefois, comme le souligne Adam dans l'extrait suivant, la cohérence est avant tout une activité interprétative : ce qui signifie qu'elle peut différer selon la lecture que l'on veut en faire :

> La cohérence n'est pas une propriété linguistique des énoncés, mais le produit d'une activité interprétative. L'interprétant prête a priori sens et significations aux énoncés et ne formule généralement un jugement d'incohérence qu'en tout dernier ressort. Le jugement de cohérence est rendu possible par la découverte d'(au moins) une visée illocutoire du texte ou de la séquence, visée qui permet d'établir des liens entre des énoncés manquant éventuellement de connexité et/ou de cohésion et/ou de progression. (Adam 1993)

A la lumière des arguments présentés ci-dessus au sujet de l'inclusion des notions de cohésion et de cohérence dans l'analyse d'un texte, nous choisissons d'écarter la quasi-totalité des éléments présentés pour ne retenir que les notions relevant d'éléments identifiables de manière objective : le raisonnement principal derrière ce choix est tout simplement d'éviter d'avoir des éléments qui recouvrent une trop grande part de subjectivité dans leur identification. Cela étant dit, les problèmes retenus pour la suite de notre analyse concernent d'une part les erreurs tout d'abord d'ordre grammatical : à savoir les erreurs de référence (et dans une moindre mesure de substitution), les erreurs de conjonctions et celles d'autre part qui relèvent d'un ordre plus sémantique en termes de progression proprement temporelle et progression thématique/informationnelle[35].

2.3.3 Vers une compétence textuelle

Alors que nous reconnaissons le fait qu'écrire en langue étrangère pose des problèmes qui ne se limitent pas à la grammaire, stricto sensu, nous estimons que la question de l'acceptation d'une production textuelle comme étant réussie dépend d'une compétence que l'on pourrait diviser en

[35] Des exemples concrets des différents types d'erreurs de progression sont fournis dans les chapitres V et VI.

deux de la façon suivante : *connaissances (proprement) linguistiques et savoir-faire fonctionnel*. Mais s'agit-il véritablement d'une nouvelle compétence à acquérir à part entière, du même ordre qu'une compétence linguistique ? Une 'nouvelle' compétence que l'on pourrait appeler *textuelle* ? Avant de tenter de définir ou plutôt de délimiter cette *compétence*, examinons quelques définitions qui nous ont précédé.

> […] le concept de compétence textuelle prend tout son sens dans la pédagogie des langues : il suppose l'appropriation de macrostructures discursives inscrites dans les règles culturelles d'interlocution ; il permet de tenir compte des « zones de savoir et d'expérience » partagées […] ; il rappelle que la fonction du langage, ce n'est pas seulement de communiquer, mais aussi de représenter l'univers qui nous entoure. (Peytard & Moirand, 1992 : 52)

Cet argument s'inscrit dans une logique analogue à celle de Kaplan (1966) voire même de Bourdieu (1982)[36], dans la mesure où le texte traduit un mode culturel d'appréhension, de représentation et d'organisation du réel. Ceci revient à suggérer que maîtriser l'expression écrite se réduit à la simple prise en compte des structures employées dans la langue étrangère et à leur reproduction. Force est de constater, cependant, que ce raisonnement n'est pas complètement recevable puisque la simple reconnaissance de la macrostructure d'un texte ne permet pas à l'apprenant-scripteur, dans aucun cas, de créer un texte similaire. En effet, les micro-constituants et les éléments de liage ne sont pas mentionnés dans cette première esquisse de définition et nous estimons donc que l'on n'y accorde qu'une faible importance. Or, comme nous le verrons plus tard, ces éléments demandent à être également appris et maîtrisés en langue étrangère, ne serait-ce que pour permettre le bon « tissage » des énoncés constituant l'objet texte.

Examinons maintenant la proposition de Charolles qui avance le propos suivant :

> Comme tout tas de mots ne donne pas une phrase, tout tas de phrases ne forme pas un texte. A l'échelle du texte ainsi qu'au plan de la phrase, il existe donc des critères efficients de bonne formation instituant une norme minimale de composition textuelle […] Ce système de règles de base (*système implicite de règles intériorisées également disponibles chez tous les membres d'une communauté linguistique*) constitue la compétence textuelle. (1978 : 8)

[36] Nous renvoyons ici à la notion d'habitus linguistique préconisé dans les travaux de sociolinguistique. (*Cf.* par exemple Bourdieu 1982 :14.)

Nous voyons ici l'introduction de la notion de règles ou critères de bonne formation[37] qui renvoient à la notion de *connexité* entre énoncés dans le discours. Ceci pourrait dès lors être considéré comme un critère textuel minimal. Au contraire donc de Peytard & Moirand, Charolles s'intéresse au fonctionnement interne du texte, c'est-à-dire, ce qui permet au texte de fonctionner et par conséquent d'exister.

Quant à Bachman (1991), il propose une taxonomie novatrice en incorporant les deux définitions ci-dessus dans ce qu'il désigne comme la compétence langagière. En effet, ce dernier regroupe les différents « composants » langagiers selon qu'ils sont employés ou associés dans le contrôle de la structure formelle de la langue, en signalant notamment les relations qu'entretiennent les signes avec leurs référents, ou employés dans le cadre d'un contexte de communication donné et selon des conventions établies. Ceux-ci sont hiérarchisés en deux groupes : une compétence dite organisationnelle pour la première et pragmatique pour la deuxième. L'aptitude à bien utiliser une langue, qu'elle soit maternelle ou étrangère, relèverait donc d'une certaine manière des mêmes compétences et connaissances. Il s'agira donc pour l'acquisition d'une langue étrangère d'un simple processus de reconfiguration. Considérons maintenant la figure 1 et ses catégorisations singulières.

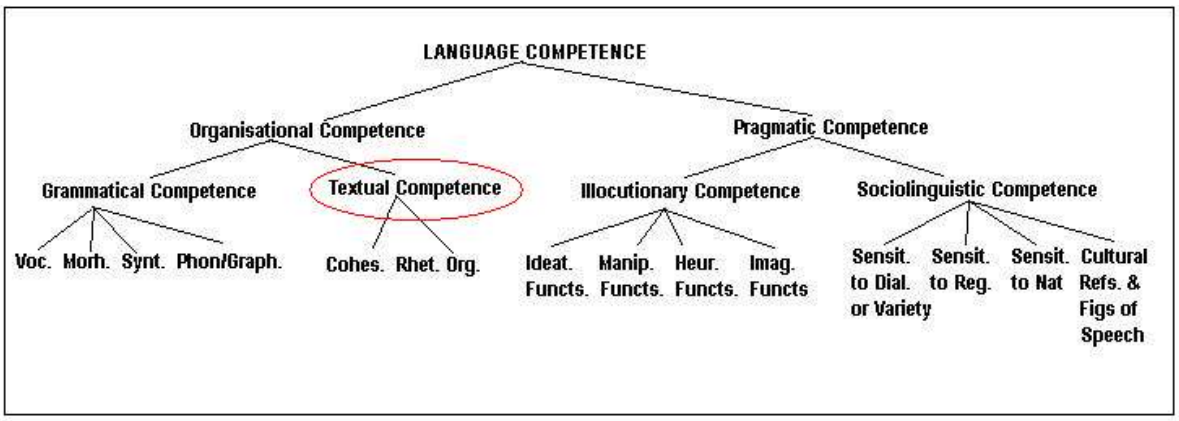


Figure 1: Language Competence de Lyle Bachman (1990)

Bachman, qui a été l'un des premiers dans la littérature anglo-saxonne à identifier et préconiser l'acquisition de la *textualité* en langue étrangère comme quelque chose de distinct et à l'inclure dans un schéma global qu'il nomme la compétence langagière, déclare :

> La compétence textuelle suppose une prise de conscience des conventions employées qui permettent de relier des énoncés ensemble créant ainsi un texte […] composé d'au

[37] *Cf*. la section précédente

moins deux énoncés ou phrases qui sont structurés selon les règles de cohésion et d'organisation rhétorique (1990 : 88)

Alors qu'il rejoint le concept clé de *liage* et de *schéma discursif culturellement façonné*, il ajoute un critère que l'on pourrait traduire par « prise de conscience ou maîtrise des conventions » et « des différentes typologies » de textes. Dans un premier temps Bachman souligne l'importance du liage au sens d'Adam (2003) avec le terme de cohésion qui, selon lui, permet non seulement de (re)lier les énoncés entre eux mais également d'indiquer les différentes fonctions cohésives qu'ils occupent dans le texte, à savoir selon la tradition d'Halliday (1976). De plus, la structure rhétorique y retrouve un rôle important au même titre que la cohésion dans la recevabilité du produit texte.

Enfin, quant à Heribert Rûck (1991), parler de compétence textuelle, revient simplement à parler d'une compétence en communication. En effet, ce dernier préconise une distinction opératoire entre ladite compétence selon qu'elle est entendue au sens large ou au sens étroit.

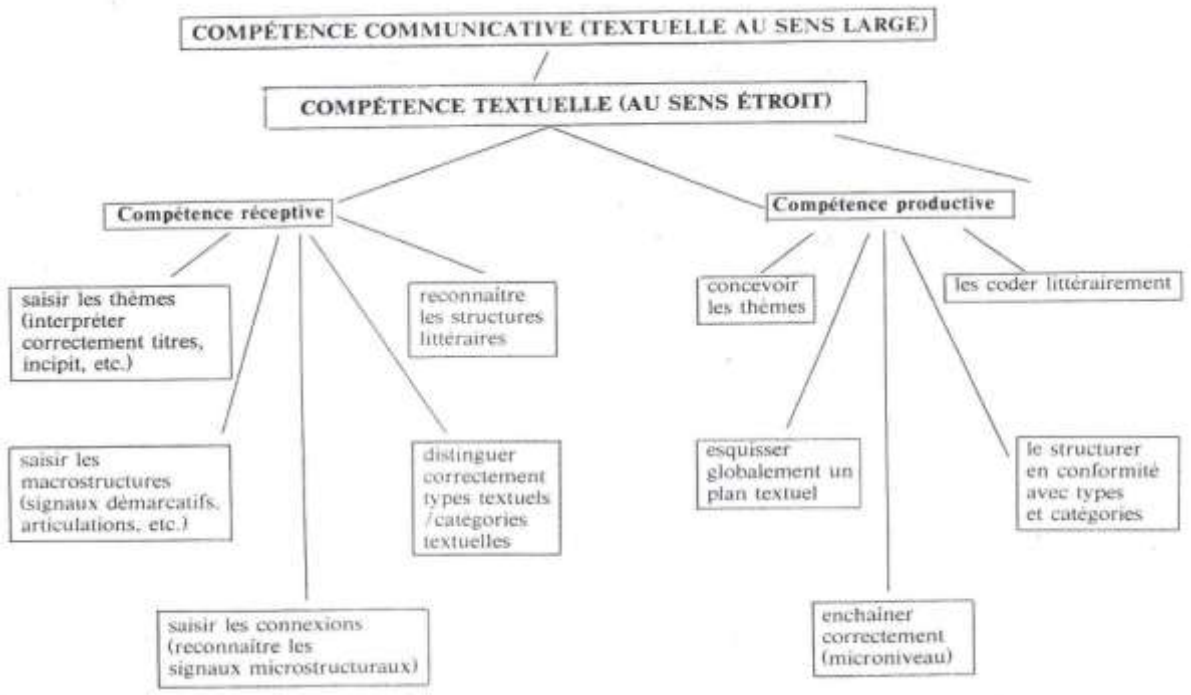


Figure 2 : La compétence textuelle de Heribert Rück (1991)

Pour reprendre le terme de Rück, cette hiérarchie *compétentielle* a le mérite d'avoir mis en exergue plusieurs micro et macro-constituants, tout d'abord en les séparant en type « réceptifs » ou productifs. Nous nous intéressons uniquement aux derniers types dans le présent travail. En effet, les arguments de Rück rejoignent ceux mentionnés ci-dessus en les regroupant ensemble de manière globale, sans pour autant donner une définition précise de ce qu'est la *compétence textuelle*. De plus, son schéma souligne l'aspect complémentaire de cette compétence dans la

mesure où celle-ci puiserait sa nature et sa fonction dans l'activité qu'elle incarne, à savoir celle d'une interdépendance pluridimensionnelle.

Au vu de ces esquisses de définitions que l'on vient d'examiner, on en déduit que les notions de bases de l'opération de *textualisation* se regroupent de la manière suivante : (i) une pratique témoignant d'une certaine maturité syntaxique et lexicale ; (ii) avec, bien entendu, l'appui d'un certain savoir-faire dans le processus de « tissage » des énoncés, opéré tout au long du texte ; (iii) à cela s'ajoute enfin une connaissance globale des stratégies ou pratiques d'écriture d'une culture discursive donnée. Or ces trois prérequis, qui demeurent des descriptions textuelles de surface, ne précisent pas la nature exacte des macro- et micro-constituants qui permettent objectivement de qualifier un écrit de texte. Cela étant dit, en partant du principe que la textualité ne relève pas strictement d'une compétence en soi à (ré)apprendre en langue étrangère, nous soutenons que celle-ci est le résultat direct, non seulement des connaissances acquises (qu'elles soient implicites ou explicites) par l'apprenant-scripteur, mais qu'elle témoigne également d'une rencontre entre conventions institutionnelles et pratiques rédactionnelles continues.

L'ensemble de ces compétences individuelles permet de développer une certaine « maturité textuelle » qui est le résultant direct des *connaissances générales* et de *praxis*. Le premier étant le niveau de base qui comprend les connaissances globales nécessaires pour commencer la rédaction d'un texte en langue étrangère, à savoir, la maturité syntaxique qu'aurait un apprenant en termes de syntaxe et de lexique. À ce niveau un apprenant saura réaliser l'unité minimale d'un texte, c'est-à-dire, des phrases individuelles, dépourvues d'erreurs majeures. Le schéma discursif culturel doit également être connu afin de pouvoir reproduire un texte approprié selon les attentes culturelles[38]. Viennent ensuite les reconnaissances des différents genres textuels (et le lexique correspondant) que l'on cherche à reproduire aussi bien que les conventions qui les gouvernent.

Or dès qu'il s'agit de réaliser une suite de phrases dont l'ensemble porte et crée un message construit, d'autres connaissances doivent être mobilisées, et ce sont celles que nous appelons des savoir-faire macro-textuels. En effet, ce qui permet de parfaire le texte comme un objet complet et uni serait la capacité ou le savoir-faire qu'aurait l'apprenant-scripteur à manier ces connaissances au profit d'un agencement hiérarchisé reflétant aussi bien sa maîtrise linguistique que discursive. Cette maîtrise ne serait évidente que lorsque ce dernier ferait apparaitre une véritable valeur argumentative personnelle.

[38] Nous rappelons que nous abordons la textualité en langue étrangère et soulignons que l'acceptabilité d'un texte n'est pas arbitraire mais varie selon les conventions textuelles d'une langue à une autre.

Nous assumons donc que cette « compétence » ne serait pas à (ré)acquérir – à proprement parler – en langue étrangère, mais nous soulignons qu'il s'agirait plutôt de stratégies textuelles (ou ce que l'on pourrait appeler rédactionnelles) qui devront être réajustées, telle la façon de structurer l'information dans le texte et la façon globale d'encadrer le discours. De surcroit, séparer ainsi la compétence textuelle en connaissances générales et savoir-faire fonctionnel nous permet de mieux situer et identifier le niveau de dysfonctionnement textuel relevé dans les écrits de langue étrangère. Nous reviendrons sur cette compétence dans le chapitre VI dans lequel des exemples concrets d'erreurs d'acceptabilité textuelle sont fournis.

(Chapitre III) L'apport de la linguistique systémique fonctionnelle

Dans ce chapitre, nous introduisons la théorie de la linguistique systémique fonctionnelle et nous précisons ce que ce cadre apporte de spécifique à notre travail. Tout d'abord, nous décrivons la théorie brièvement afin de la distinguer de la grammaire traditionnelle, avant de nous prononcer sur les principaux éléments qui ont été développés dans ce modèle que nous appelons systémique. A cette fin, nous précisons les éléments permettant d'établir le cadre théorique qui sera employé dans l'analyse de toute unité jugée erronée, telles qu'elles ont été décrites dans les deux chapitres précédents, autrement dit, tantôt les unités lexicales individuelles, tantôt les unités multi-mots ou leurs équivalents appartenant à une catégorie linguistique supérieure.

Du fait de l'aspect original de l'approche, il sied de rappeler les contours théoriques de la linguistique systémique fonctionnelle qui fournit une véritable architecture de la langue, de manière à mieux comprendre l'une de ses particularités appelée *les métafonctions sémantiques*. Ces métafonctions seront discutées en détail, car elles constituent à la fois le point de départ et le point d'analyse de la présente étude.

3.1 Qu'est-ce que la linguistique systémique fonctionnelle (LSF) ?

Définir ou décrire la complexité de la langue en tant qu'activité humaine n'est pas chose aisée ; il en va de même si l'on cherche à définir une théorie du langage. Le piège étant d'être soit trop simpliste, soit trop prolixe. Néanmoins, nous allons tenter de décrire les fondements de la théorie de la linguistique systémique fonctionnelle (LSF) afin de fournir une base cohérente et solide pour ce qui est à venir. Dans cet esprit, nous retracerons assez brièvement les débuts conceptuels de la

linguistique systémique afin de mieux situer la théorie, telle qu'elle s'est développée aujourd'hui. De plus, nous aborderons les principales sources d'inspiration de certains concepts en LSF : (i) de manière à distinguer ceux qui sont proprement systémiques (cf. les métafonctions) de ceux qui sont partagés entre plusieurs courants dits structuralistes ou fonctionnalistes et (ii) afin de comprendre l'étendue de leurs utilisations en LSF – souvent avec des différences notables par rapport au concept original (cf. par exemple l'axe paradigmatique).

Mais avant d'aller plus loin, nous pouvons d'ores et déjà soutenir que la théorie dite de la linguistique systémique fonctionnelle se veut une théorie sémiotico-sociale : *sémiotique* parce que l'on s'intéresse avant tout à la construction du sens ou à la signification comme résultant d'un processus de composition et de contextualisation ; et *social* puisque la langue est appréhendée foncièrement dans sa dimension d'usage où son emploi est régi et répond à la fois à des codes et des besoins sociaux. A cela s'ajoute bien entendu le fait que la langue en LSF désigne un système de potentiel corrélé, c'est-à-dire établi en système de réseaux de niveaux différents se réalisant tous de manière simultanée.

Le fait donc de choisir un item linguistique par rapport à un autre (qu'il soit individuel ou composé) est, de ce fait, considéré comme porteur de sens. Dans cette conception, le choix des mots, des syntagmes, des propositions à travers les différentes typologies phrastiques (déclarative, interrogative ; simple, complexe …) indique la valeur sémantique précise à attribuer aux différentes sélections. Il en est de même d'un regard sur le système en considérant les choix qui ont été opérés, ainsi que ceux qui ont été exclus, puisque ceci permet également de comprendre les limites de la valeur interprétative autorisée par le locuteur/scripteur au destinataire. Et c'est justement tous ces aspects multifonctionnels voire multi-compositionnels qui nous intéressent dans notre étude – puisqu'ils facilitent la description des différentes erreurs relevées dans notre corpus, en apportant des étiquetages divers et variés permettant ainsi d'accéder à une perspective qui, jusque-là, a été très peu développée.

3.1.1 L'origine de la théorie systémique

Toutefois, comme nous l'avons soutenu ci-dessus, définir une théorie n'est pas chose aisée, ainsi faut-il examiner les concepts clés de la théorie pour avoir une meilleure compréhension globale du modèle théorique mis en jeu. A cet égard, nous soulignons que la LSF trouve ses origines dans les travaux de Michael Halliday à partir des années 1950, quand ce dernier s'intéressait au système linguistique chinois – à la fois en tant qu'objet d'étude et objet pédagogique. Cette conjoncture

doublement didactique et théorique constituait donc le point de départ de ses premières descriptions linguistiques. Mais, comme nous le verrons plus en détail ci-dessous, malgré le fait que la linguistique systémique fonctionnelle présente aujourd'hui des spécificités singulières, il s'agit plutôt d'une théorie évolutionnaire et non d'une théorie dite révolutionnaire en ce qu'Halliday s'est fortement appuyé sur les concepts fondés par ses prédécesseurs pour ensuite identifier et développer davantage les éléments jugés complémentaires (Matthiessen : 2007). Mais l'élément déclencheur principal – à proprement parler – demeure selon Gonzaga (2011) la première rencontre entre réflexions linguistiques d'un côté et pratiques didactiques de l'autre. Et cette rencontre a provoqué une prise de conscience du phénomène qui est devenu par la suite le socle fondamental de la théorie systémique.

> When Halliday taught his first Chinese class in 1945, he noticed that the clause should be the "centre of action in the grammar" (Halliday, 2009, p.355). At that time the clause was not acknowledged as "a general organizing category" (ibid). It was not "the locus, where fundamental choices in meaning were acted out" (ibid). We might say that this observation led Halliday to treat grammar in a different fashion. (2011 : 34).

Cette prise de conscience a indéniablement été le moteur derrière la première version de la théorie appelée initialement *Scale and Category Grammar (SCG)*[39]. Cette version qui n'a eu qu'une « courte vie » a eu un impact considérable sur la linguistique de l'époque (Fawcett 2000). De plus, elle a introduit deux notions clés qui ont illustré le besoin naissant de séparer les différents niveaux d'analyse selon qu'ils relèveraient de ce que l'on associe a priori aujourd'hui aux axes dits syntagmatique et paradigmatique. Ces deux notions clés se traduisaient par (i) *Grammatical categories* et (ii) *Scales,* toutes deux sous-divisées respectivement en « *unit, structure, class, system* » et « *rank, exponence, delicacy* ». Outre ces deux notions, cette première version a posé la pierre angulaire privilégiant la séparation en niveaux distincts selon l'élément linguistique à l'étude (cf. par exemple la stratification 3.2.2). Cet aspect est clairement indiqué à travers ces brèves descriptions mises en avant par deux auteurs :

> The theory saw the linguistic system as comprising the level of form, itself made up of lexis and grammar, together with two interlevels, context and phonology. Context provided the link between situation and form, and phonology provided the link between form and sound. The grammar set out to handle the analysis of stretches of language

[39] Eu égard au fait que certaines de ces terminologies ont été abandonnées ou que leur sens de départ a évolué, nous trouvons judicieux de ne pas les traduire ici afin de ne pas les confondre avec d'autres termes propres à la théorie LSF et qui seront explicités plus tard dans ce chapitre. Pour donner un exemple d'un abandon terminologique, *Categories* renvoyait aux deux niveaux distincts de grammaire et lexique, avant que ces derniers ne soient reclassés au même rang dans la version actuelle. *Cf.* section 3.2.

> that had actually occurred, so that at this stage the grammar was descriptive rather than generative. (Morley 2000 : 1)
>
> However its major focus was on 'grammar', in a sense that is roughly equivalent to a combination of the traditional senses of the terms "syntax" and "morphology"[...] The central concept is that of levels of language. Halliday states that "the theory requires that linguistic events should be accounted for at a number of different levels", and he then goes on to claim that "the primary levels are of 'form', 'substance' and 'context'". (Fawcett 2000 :17)

L'évolution donc de la *Scale and Category Grammar* vers la grammaire systémique fonctionnelle[40] (GSF) et ensuite la Linguistique Systémique Fonctionnelle (LSF) résulte de la rencontre entre plusieurs courants linguistiques, allant notamment de l'incorporation des fondements du professeur de Halliday, John Rupert Firth et de son collègue Bronislaw Malinowski à la pensée fonctionnaliste, issue du cercle linguistique de Prague. Notons cependant que ces trois influences ne sont que les plus communément admises et attestées, et ne peuvent en aucun cas justifier la transition intégrale de la SCG à la LSF. Mais ce qui est non-négligeable dans cette nouvelle désignation est l'accent mis sur l'héritage d'une part *systémique* de Firth et d'autre part *fonctionnelle* du cercle de Prague. Il est également à souligner que Bloor &Bloor (2004) dressent une liste non-exhaustive des courants et rencontres les plus influents ayant contribué au développement de la théorie. Nous avons schématisé, à titre illustratif, l'ensemble de ces « rencontres » dans la figure 3, page 69.

Un tel schéma n'a certes pas de prétention à l'exhaustivité mais permet aisément de situer les principales influences conceptuelles réunies dans la théorie hallidayienne. Nous notons cependant par exemple que, si l'influence de Firth, Malinowski et du cercle linguistique de Prague est traditionnellement acceptée (cf. Bloor & Bloor 2004 ; Matthiessen 2007 ; Hasan 2009), celle de Bühler et Hjelmslev le sont dans une moindre mesure (cf. Taverniers 2011 ; Bache 2010). En outre, le graphique ci-dessous met en avant l'origine de quatre notions essentielles à la LSF actuelle : à savoir (i) *la stratification,* qui permet de diviser la grammaire en plusieurs strates distinctes : par exemple avec la phonologie, la lexicogrammaire et la sémantique (Halliday & Matthiessen 2004 : 34) ; (ii) *la fonction,* qui signifie que « priority is given to the view 'from above'; that is, grammar is seen as a resource for making meaning » (ibid.) ; (iii) *le système,* qui renvoie à la fois au fait que la grammaire est entendue comme « a network of interrelated meaningful choices » par le biais de « system network, not as an inventory of structures » (ibid.) ; et enfin (iv) *le contexte,* qui voudrait

[40] L'appellation « *Systemic Grammar* ou *Functional Grammar* », employée souvent de manière interchangeable, est apparue moins de 10 ans après la publication de ce qui dessinait le contour de *Scale and Category Grammar*.

que le contexte d'énonciation, tant situationnel que culturel, soit pris en compte dans toute analyse systémique. Ces quatre notions seront explicitées plus en détail ultérieurement.

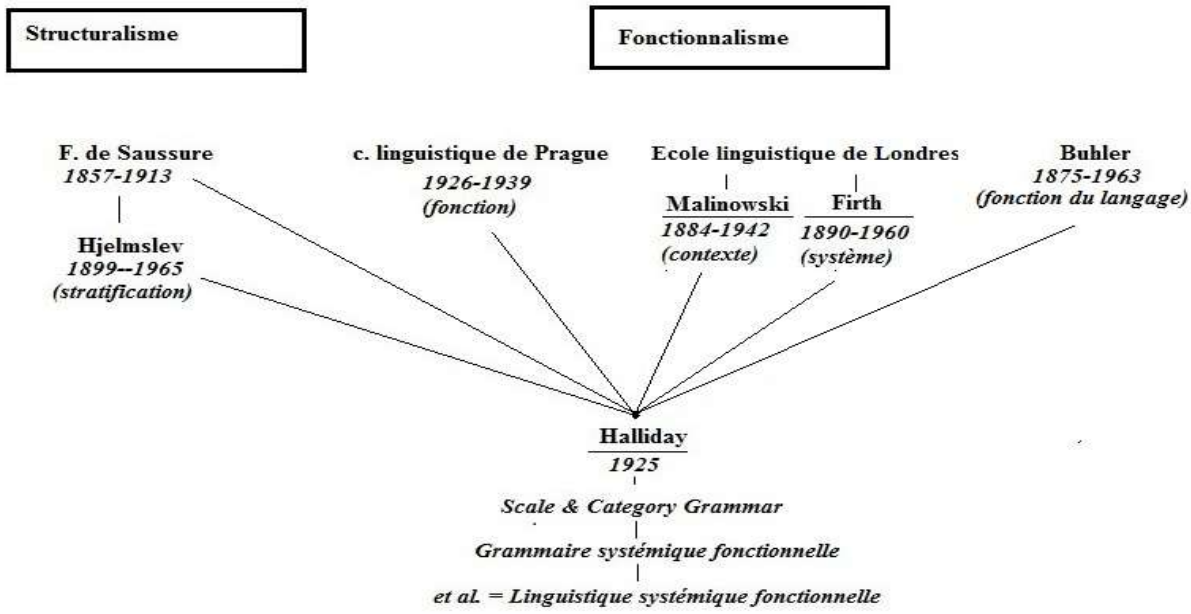


Figure 3 : Les influences principales de la linguistique systémique fonctionnelle

Une fois les influences situées, nous pouvons désormais préciser que la théorie systémique s'oppose à ce que l'on désigne communément sous l'étiquette de grammaire traditionnelle et tout singulièrement de grammaire générative[41]. La première renvoyant à l'enseignement grammatical généralement dispensé à l'école et la seconde à la théorie du langage développée par Noam Chomsky. La distanciation de LSF par rapport à ces deux traditions réside dans l'acceptation globale de la langue, entendue principalement comme un outil de communication socialement ancré. Cette dimension sociale se retrouve dans deux notions clés : à savoir le contexte situationnel et le contexte culturel (de Firth et Malinowski). Ces deux notions se traduisent par l'intérêt accordé au processus de construction globale du sens, de même que les choix effectués et l'implication de ces choix sur le sens précis qui en découle. En d'autres termes, la LSF s'intéresse à l'*utilisation effective* et *réelle* de la langue, tant d'un point de vue paradigmatique que syntagmatique, et non à l'approche prédictive et prescriptive de l'ensemble des combinaisons grammaticalement possibles.

[41] Cf. Gledhill (2011) pour une comparaison entre les deux modèles. Voir aussi à titre d'information Butler (2003) pour une comparaison entre trois modèles linguistiques dits fonctionnels : LSF, la *Role and Reference Grammar* et *Functional Grammar*.

3.1.2 L'application de la théorie

Dès lors que les influences majeures de la théorie sont identifiées, on est à même de comprendre le raisonnement derrière la méthode d'analyse. Il nous est désormais possible de soutenir que la théorie systémique fonctionnelle peut être assimilée à une structure de systèmes stratifiés et interdépendants.

Ces notions d'interdépendance et de stratification sont centrales à la théorie, telle qu'elle a été développée par Halliday depuis les années 1960. En effet, pour aller plus loin dans cette perspective il est important de concevoir la double référence « systémique fonctionnelle » comme renvoyant d'abord à une conception de la langue dans laquelle celle-ci est envisagée en tant que système sémiotique utilisé principalement pour créer du sens dans un contexte social précis. Tandis que l'aspect proprement fonctionnel permet d'identifier les unités linguistiques mobilisées dans cette création de sens et de porter un regard, telle une radiographie, sur les unités constituantes (l'axe syntagmatique). De plus, cette nouvelle optique faciliterait la description systématique. Prenons par exemple le tableau suivant qui fournit une vision d'ensemble sur un énoncé court.

Sémantique	system de métafonctions réalisé par la *phrase*																						
Lexicogrammaire	D	SN		SV			S.Adj					SN		S.P	D	S.Adj			SN				
	The	**cartoon**		**underlines**			**environmental**					**issues**		**in**	**our**	**globalised**			**economy**				
Graphologie	The	car	toon	un	der	lines	en	vir	on	men	tal	is	sues	in	our	glo	bal	lised	e	con	o	my	
Phonologie	*thə*	*kär*	*tōōn*	*ŭn'*	*'dər*	*līn'*	*ĕn*	*vī'*	*rən*	*mĕn'*	*tl*	*'ĭf*	*ju:*	*ĭn*	*avə*	*glō'*	*bə-*	*līzd*	*ĭ-*	*kŏn*	*ə-*	*mē*	

Tableau 5 : Un aperçu de la conceptualisation de la théorie systémique

Cette description résolument simple a le mérite de schématiser la théorie systémique avec son dénominateur le plus large. Il n'est donc pas possible de « dézoomer » davantage, mais son contraire est tout à fait possible. Le plan sémantique peut par exemple se diviser en trois sous-parties selon la valeur que l'on cherche à étudier. De ce point de vue, la LSF se veut alors une théorie extensible et circonstanciée - en ce qu'elle propose d'examiner les différents aspects de la langue selon leur contexte d'utilisation et le sens précis qui en découle, tout en portant une attention particulière aux choix [linguistiques] qui ont été opérés.

Reprenons l'exemple de l'énoncé ci-dessus. A titre indicatif, celui-ci peut être examiné selon ses différentes valeurs contextuelles. C'est-à-dire qu'une analyse sémantique fonctionnelle donnera le tableau suivant :

TRANSITIVITE	Porteur	Procès (relationnel)	attribut	circonstance locative			
MODE	Mode		Résidu				
	sujet	conjugue/ prédicat	complément	ajout			
THEME	Thème	Rhème					
	The cartoon	underlines	environmental issues	in	our	globalised	economy

Tableau 6 : Un aperçu de la strate sémantique

Ici, un apport de la LSF par rapport à la grammaire traditionnelle réside dans la mise en évidence de la Transitivité[42], de sorte à préciser à la fois la valeur sémantique et systémique des choix lexicaux qui ont été faits : ce qui n'est pas anodin. Ces choix témoignent, entre autres, de la façon donc le locuteur conçoit son message – à travers, par exemple, l'intensité factuelle ou subjective qu'il souhaite y apporter. Cela étant, chaque élément de l'énoncé apporte une information qui dépasse le cadre syntaxique du *sujet, verbe, objet* dit SVO, ou encore *acteur, verbe, patient*. Le présent modèle a été repensé de manière à apporter ces précisions sémantiques sur le rôle que jouent les éléments individuels par rapport au noyau verbal, qui devient désormais un élément clé (*cf.* sections 3.2.5 et 3.4).

Il est également possible de s'intéresser à ce même énoncé selon sa construction grammaticale, en termes de structuration par rapport à la langue en tant que système. Ce qui prime dans la figure 4 serait alors la hiérarchisation et le rapport d'interdépendance des éléments les uns par rapport aux autres.

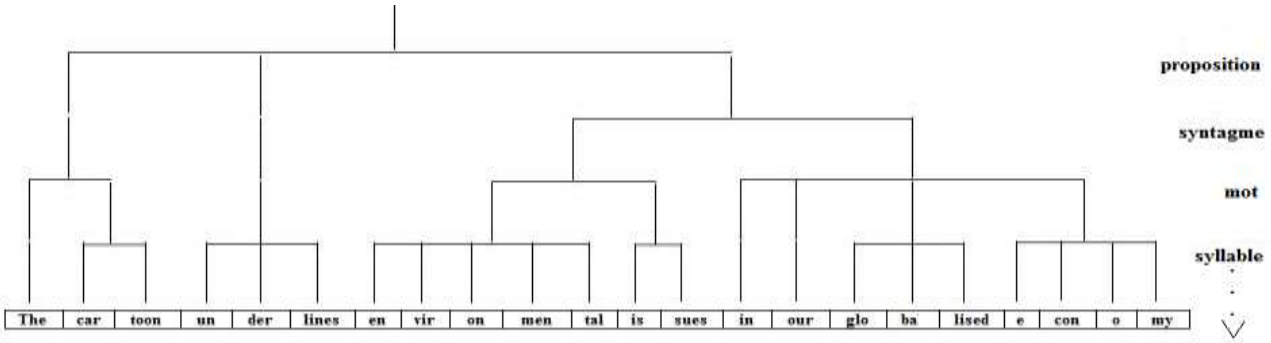


Figure 4 : La schématisation grammaticale

En plus des deux premières notions d'interdépendance et de stratification, ces trois premiers graphiques mettent en exergue un principe tout aussi fondamental de la théorie LSF ; à savoir, la contextualisation. Cela signifie que la construction du sens ne résulte pas de la simple addition des unités syntaxiques, mais que le contexte y participe de manière prépondérante. A titre d'exemple,

[42] *Cf.* section 3.2.5 pour une présentation détaillée.

toute analyse faisant appel à la théorie systémique doit être réalisée en tenant compte du contexte textuel environnant de l'unité à l'étude.

Et c'est justement en raison de ces trois concepts clés que la LSF fait appel à divers schémas sémiotiques pour représenter les différents types ou niveaux d'analyses effectuées. En bref, ces schématisations ne sont pas rares chez les analystes systémiques, en raison du fait qu'elles fournissent un visuel non-négligeable sur les unités linguistiques minimales à l'étude ainsi que pour l'ensemble des occurrences linguistiques insérées dans un niveau d'analyse précis (cf. *les strates*). Toutes ces schématisations ont pour but d'illustrer le niveau de corrélation par rapport au système global.

3.2 Le modèle architectural de la langue en LSF

Si l'on souhaite s'intéresser davantage à la linguistique systémique fonctionnelle, il est important que le cadre conceptuel global soit explicité de manière articulée. Ce cadre renvoie, ici, de façon générale à ce que Halliday a sciemment dénommé l'*architecture de la langue* (2003, 2004). Et bien que cette désignation ne signale pas un nouveau point de vue théorique, elle a pour but de préciser les principes organisateurs déjà existants dans ses publications antérieures, mais présentés dorénavant de façon regroupée et moins disparate. Cette « mise à jour » a donc le mérite de rendre le modèle plus compréhensible en ce qu'elle adopte l'approche du plus général vers le spécifique : autrement dit, en commençant par les principes organisateurs du système langagier avant d'aborder les descriptions théoriques délicates fondées sur des considérations systémiques spécifiques.

Halliday explique ce besoin de réorganisation en soulignant que : « *anyone coming to read these chapters* [comprendre, les ouvrages] *is entitled to ask what sorts of things about language are being taken for granted – and what even more, perhaps, what things are not being taken for granted* ». (2003 : 3) Dans cette optique, nous allons tenter de présenter succinctement le 'modèle architectural' afin de poursuivre notre bref aperçu d'introduction entrepris dans la section 3.1.

3.2.1 La stratification

Comme nous l'avons soutenu dans la section 3.1.2, le concept de *stratification* occupe une place indispensable dans la théorie systémique. Il renvoie singulièrement au principe d'organisateur global permettant à la fois de contextualiser, d'ordonner et de regrouper les différents types d'analyse selon la strate du système linguistique à l'étude. Ceci peut être appréhendé au moyen du cas de figure suivant.

Avec des apprenants de langue étrangère, disons que nous voulons étudier leur production orale et écrite. Pour l'oral, il pourrait être question d'étudier la production de segments sonores de manière globale ou de prêter une attention particulière à la segmentation de sons discrets et distincts présents dans la langue cible mais non encore maitrisés chez les apprenants. Autrement dit, la prononciation de certaines voyelles et consonnes ou le rythme ou intonation globale adoptés. Inutile de dire ici que pour la production écrite, la grammaticalité et l'orthographie pourraient être parmi les premiers éléments étudiés : et ce, avant de s'aventurer sur des aspects plus délicats tels que l'acceptabilité par rapport au contexte de rédaction ou le genre textuel attendu. La stratification en LSF permet donc de visualiser ces différents niveaux d'analyse linguistique en montrant les relations directes ou immédiates entre les éléments étudiés par rapport à l'ensemble du système de la langue.

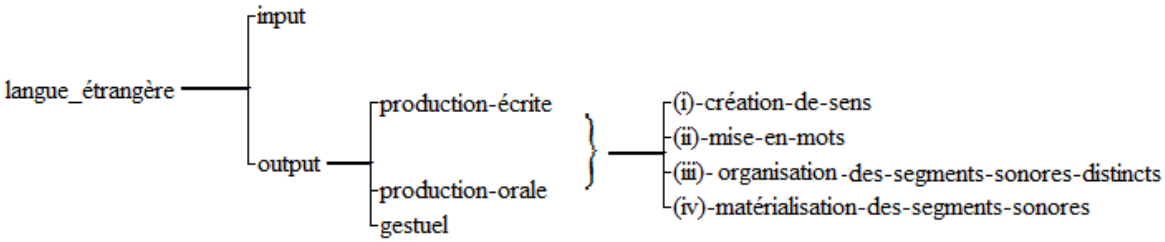


Figure 5 : Un exemple de stratification dans la production en L2

Cette représentation graphique permet de faire un premier constat lors de la production du discours : à savoir, celle de la dépendance des éléments en fin de l'*axe output* (i-iv) les uns par rapport aux autres dans la création du sens final. En effet, nous pouvons voir que le sens final de l'objet à l'étude résulte d'abord de la *mise en mots*, qui à l'oral se réalise par le biais de segments sonores distincts et ainsi de suite. En LSF, ces quatre instances en fin de l'axe *output* représentent chacun un niveau d'analyse différent ; de ce fait, ils comportent chacun une valeur sémiotique intrinsèque.

i. Créations de sens	Sémantique	Contenu
ii. Mise en mots	Lexico-grammaire	
iii. Organisation des segments sonores distincts	Phonologie	Expression
iv. Matérialisation des segments sonores	Phonétique	

Tableau 7 : La stratification approfondie

Pour aller plus loin dans cette analyse, la stratification permet de distinguer toute utilisation de la langue en fonction du cadre fonctionnel mis en jeu : soit en termes de contenu (comprendre, le sens final) soit en termes de l'expression des moyens employés pour réaliser le *contenu*. Suite à cela, les éléments sont ordonnés selon l'environnement linguistique immédiat par rapport à la langue en tant que système dans tous les emplois sémiotiques que ce soit.

Au vu de la complexité de la langue, la stratification regroupe donc les éléments selon un principe d'interdépendance ou leur lien de corrélation. Halliday & Matthiessen (2004) proposent donc de visualiser la langue en ce qu'ils nomment les cinq dimensions de la langue (*cf.* tableau 8). Notons que ce visuel illustre, dans une certaine mesure, la stratification à la fois en tant que principe organisateur à une petite échelle et à l'échelle du système entier, puisque le fait de catégoriser les dimensions relève lui aussi de la stratification.

	Dimension	**Principe**	**Ordre**
1	Stratification	Réalisation	sémantique ~lexico-grammaire ~phonologie ~phonétique
2	Instanciation	actualisation	potentiel ~ sous-potentiel ou type d'instanciation ~ instance
3	structure (ordre syntagmatique)	Rang	proposition ~syntagme ~ mot ~ morphème
4	système (ordre paradigmatique)	Finesse	Grammaire ~ lexique [lexico-grammaire]
5	métafonction	métafonction	Idéationnelle ~ interpersonnelle ~textuelle

Tableau 8 : Les cinq dimensions identifiées dans la langue : adapté de Halliday & Matthiessen 2004

3.2.2 L'instanciation

L'instanciation est le processus par lequel la langue est considérée par rapport à son potentiel systémique global et son utilisation effective dans un contexte donné. Autrement dit, c'est la relation qu'entretient la langue en termes de potentiel sémantique infini et une occurrence linguistique individuelle. Il est important de souligner ici que l'instanciation est un élément clé dans la conception systémique de l'analyse du langage ; et ce, de la même manière que la stratification. Notamment en raison du fait qu'elle renvoie à des considérations métalinguistiques, permettant à l'analyste de visualiser à nouveau l'ensemble du système de la langue lors de son analyse. Mais au lieu d'ordonner les différents niveaux d'analyse en strates, le principe sous-jacent est d'établir un lien direct entre l'objet linguistique produit en fin de discours et sa relation par rapport au système global. Cette conception relationnelle a été nommée *observer perspective* par Matthiessen (2007 : 515).

De manière analogique, Halliday & Matthiessen comparent la relation entre le système de la langue et l'instanciation d'une occurrence donnée, aux systèmes météorologiques : le climat entendu au sens large serait l'ensemble du potentiel du système tandis que le temps qu'il fait à un moment donné serait l'instanciation dudit système.

> Climate and weather are not two different phenomena; rather, they are the same phenomenon seen from different standpoints of the observer. What we call 'climate' is weather seen from a greater depth of time — it is what is instantiated in the form of

> weather. The weather is the text: it is what goes on around us all the time, impacting on, and sometimes disturbing, our daily lives. The climate is the system, the potential that underlies these variable effects. (2004 : 27).

Cependant, en dépit du fait que ces deux caractéristiques de composition nommés *potentiel* et *instance* (*cf.* Figure 6) remontent au début des années 1970, il a fallu attendre les années 1980 et 1990 pour que le concept soit repris, théorisé et testé par des algorithmes en linguistique informatique (Matthiessen 2007 : 528-529). Il est également à noter que Matthiessen souligne le fait que ces deux terminologies ont été introduites de manière à désambiguïser le terme firthien initial d'*exponence*[43] qui avait été repris par Halliday lors de ses premières descriptions du phénomène et notamment dans la première version de la *Scale and Category Grammar*. (cf. section 3.1.1).

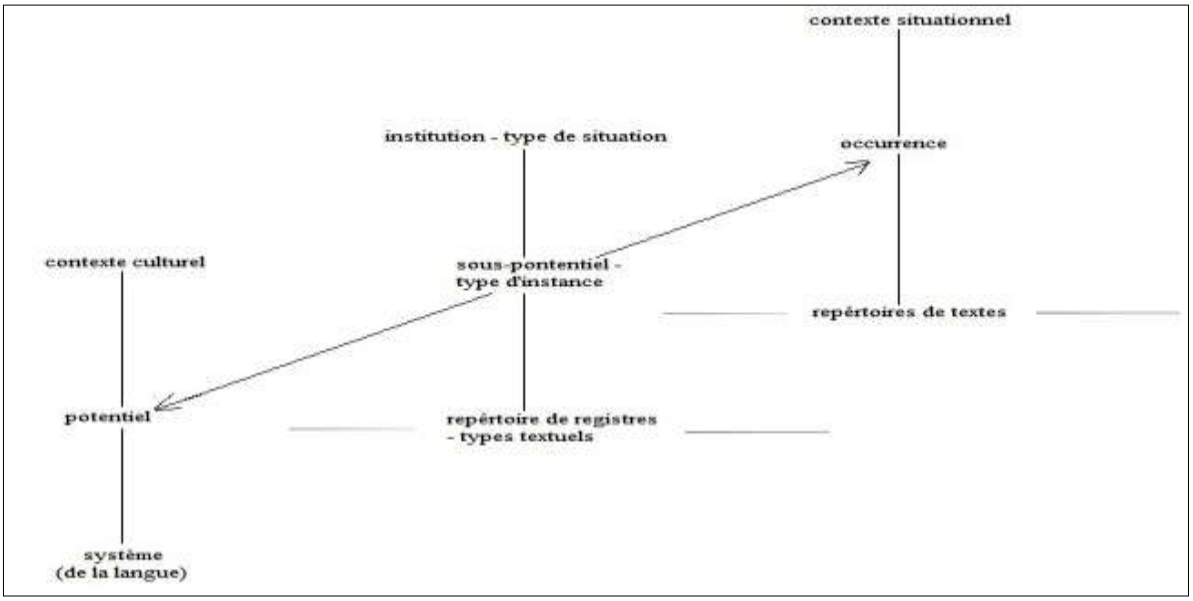


Figure 6 : L'axe d'instanciation de Halliday& Matthiessen 2004 (reprise et adaptée par Matthiessen 2007)

De plus, Halliday fournit un graphique illustrant la mise en commun à la fois des principes de l'instanciation et de la stratification. Ceci a le mérite, singulièrement dans notre projet de recherche, de constituer une nouvelle façon de classer, d'interpréter et éventuellement de remédier aux occurrences erronées de notre corpus. Et ce, en raison du fait que cette mise en commun apporte un nouveau regard sur une question épineuse : à savoir comment expliquer et classer les erreurs en langue étrangère. D'après la figure 6, nous pouvons émettre l'hypothèse que certaines erreurs seront plus près de certains points sur le graphique. Par exemple une erreur se situant près de

[43] Notons que cette clarification a permis de préciser davantage les paires « potentiel – instance » du principe de « réalisation et actualisation » (cf. tableau 8 : Les dimensions identifiées dans la langue)

« système de la langue » serait plutôt d'ordre grammatical, tandis qu'une autre se situant au-dessus du « potentiel » serait grammaticale mais non acceptable en raison du contexte.

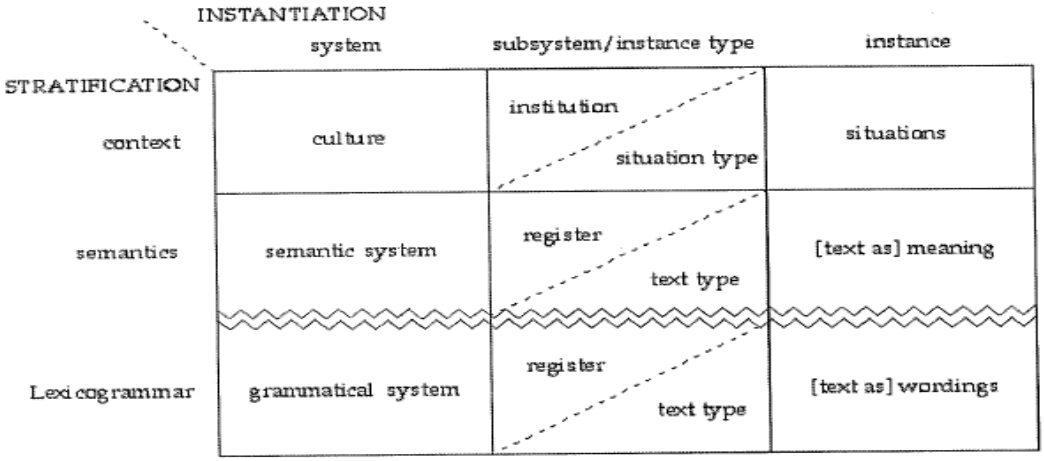


Figure 7 : Le regroupement des axes d'instanciation et de stratification (Halliday 2002, Matthiessen 2007)

En outre, en extrapolant davantage d'après la figure 7 ci-dessus, il nous est possible de soutenir que les erreurs situées au niveau lexico-grammatical seront limitées à des erreurs grammaticales ou des erreurs de lexique (dites erreurs du système de base) ; celles relevant de la construction du sens se limiteront donc à la strate sémantique ; et le contexte culturel serait responsable pour les différents types d'écart situationnels ou institutionnels. (*cf.* chapitres I et II). Ceci peut être schématisé davantage en partant de haut en bas par des erreurs d'acceptabilité, de compréhension et de grammaticalité. Notons, toutefois, que cela ne signifie pas qu'ils sont distinctement séparés les uns des autres ; mais plutôt qu'ils facilitent une meilleure compréhension de certains phénomènes linguistiques.

3.2.3 L'ordre syntagmatique (structure)

La troisième dimension hallidayienne dans l'architecture de la langue est celle de l'ordre syntagmatique[44]. Il est important de souligner ici que ce concept est très répandu suivant les différentes théories linguistiques, mais avec des considérations assez disparates. Il nous parait judicieux donc de redéfinir sa conception tant en linguistique traditionnelle qu'en linguistique systémique, de manière à le désambiguïser pour la suite. En effet, la syntagmatique est une des notions phares dans la linguistique moderne depuis son introduction dans le *cours de linguistique générale* de F. de Saussure (1979)[45]. Ce dernier y voyait trois caractéristiques propres : tout d'abord

[44] Nous adoptons le terme *ordre syntagmatique* au lieu de l'*axe syntagmatique*, en raison du fait que la valeur sémantique accordée à l'ordre fonctionnel des constituants est elle-aussi porteuse de sens dans le cadre LSF.
[45] Il s'agit de l'édition critique, éditée chez Payot. La version originale étant de 1916.

celle d'un rapport double, à savoir un enchaînement d'unité avec des rapports fondés sur le caractère linéaire de la langue (ibid. : 170) ; ensuite le fait qu'un rapport syntagmatique suppose que l'analyse se fait uniquement à partir des éléments *in praesentia* (ibid. : 171) ; et enfin le principe de compositionnalité a été esquissé.

De manière générale, la première caractéristique renvoie à l'enchainement au sens strictement syntaxique dans lequel les éléments individuels se combinent selon des règles d'agencement définies. Le tableau 5 et notamment la figure 4 témoignent de cet aspect syntaxique en identifiant quelques éléments de base avec leur combinaison au niveau des syntagmes. La notion de *in praesentia* suppose que l'analyse syntagmatique ne peut comprendre que les éléments effectivement employés dans l'énoncé – ceci est à mettre en opposition avec l'ordre paradigmatique (*cf.* section 3.3.3) qui prend en compte non seulement les éléments employés, mais également ceux qui auraient pu l'être ou encore ceux qui pourraient par exemple faire l'objet d'une commutation avec des éléments d'une même catégorie grammaticale. Enfin le principe de compositionnalité n'a pas été clairement identifié en tant que tel. Cependant, l'allusion est faite – dès lors que de Saussure souligne la chose suivante :

> La notion de syntagme s'applique non seulement aux mots, mais aux groupes de mots, aux unités complexes de toute dimension et de toute espèce (mot composés, dérivés, membres de phrase, phrases entières). Il ne suffit pas de considérer le rapport qui unit les diverses parties d'un syntagme entre elles […] il faut tenir compte aussi de celui qui relie le tout à ses parties. (1979 : 172)

Ces multiples considérations n'ont pas été reprises ou étudiées voire « adoptées » de la même manière. A titre d'exemple, certains limitent désormais le phénomène syntagmatique aux règles de combinaison syntaxiques : c'est-à-dire, la première caractéristique mise en avant par Saussure. La définition de Johnson et Johnson (1999) nous donne une illustration assez représentative de la réduction radicale que le concept a subi au cours des années. Ces derniers soutiennent, en effet, que le rapport syntagmatique se limite à l'étude des unités qui précèdent ou se succèdent sur une séquence linéaire. Or, il semblerait que de Saussure ait essayé de mettre en garde contre cette réduction en affirmant :

> […] la syntaxe, c'est-à-dire, selon la définition la plus courante, la théorie des groupements de mots, rentre dans la syntagmatique, puisque ces groupements supposent toujours au moins deux unités distribuées dans l'espace. Tous les faits de syntagmatique ne se classent pas dans la syntaxe, mais tous les faits de syntaxe appartiennent à la syntagmatique (1979 : 188)

Hélas, la syntagmatique aujourd'hui renvoie dans un premier temps aux rapports syntaxiques – ou combinatoires entre unités grammaticales dans un énoncé donné et la notion de composition s'avère reléguée au second plan. C'est justement ici que nous relevons un des premiers points de démarcation de la syntagmatique chez Halliday, dans la mesure où la composition est remise au premier plan. Dans sa conception actualisée, le principe organisateur de ce niveau d'analyse est ce qu'il nomme le « rang » (*rank*, en anglais). Le principe de compositionnalité se retrouve alors comme objet d'étude principal dans cette strate.

Dimension	**Principe**	**Ordre**
Ordre syntagmatique (structure)	Rang	proposition ~ syntagme ~ mot ~ morphème

Tableau 9: L'ordre syntagmatique (LSF)

Nous notons cependant que la conception saussurienne de la syntagmatique se démarque de celle de Halliday sur un autre point. Pour le premier, elle se limite à l'étude de l'énoncé en raison notamment du fait qu'il est entendu comme étant le dénominateur le plus large de l'analyse syntaxique ; et ce, notamment en grammaire traditionnelle. En effet comme nous l'avons souligné (*cf.* section 3.1.1 et Bloor & Bloor (2004)), la proposition est le point de départ de l'analyse systémique – ainsi, une analyse est possible en examinant sa composition avec un « regard d'un bas », c'est-à-dire les éléments individuels qui la composent ; *a contrario*, la proposition peut également être analysée comme élément singulier dans une composition plus large. Et pour Halliday, la syntagmatique suppose avant tout d'être en mesure d'expliquer les constituants de toute structure discursive qui crée du sens – aussi petit ou large que soit l'ensemble.

Ce principe de compositionnalité (qui est traduit grossièrement de *constituency* en anglais) peut être entendu comme un principe à trois facettes qui se réalisent conjointement : à savoir compositionnel, hiérarchique et structural. Compositionnel étant donné qu'une unité lexicale est composée d'une ou plusieurs lettres ; hiérarchique au vu de la perspective obtenue après analyse, par exemple « d'un haut » ou « d'un bas » ; enfin structural puisque la combinaison de plusieurs éléments crée une unité de sens – et ce sens final n'est que le résultant direct de l'ensemble structuré selon les attentes du système linguistique global.

En définitive, il est important de souligner que la notion de compositionnalité hallidayienne ne se limite ni à la strate lexico-grammaticale ni à la strate sémantique. Elle est applicable sur le plan de l'expression – tant en phonologie qu'en phonétique et ainsi de suite (cf. tableau 7). Mais pour les besoins de la présente étude, nous examinons uniquement le plan dit du *contenu* et non celui de

l'*expression*[46]. Enfin, notons également comme le soutient Halliday (1994, 2003) que l'aspect proprement syntagmatique en linguistique systémique n'est pas le plus exploité par les systémistes. Mais il faut convenir qu'il n'en est pas moins essentiel. Tout cela s'explique par le fait que la *paradigmatique* prime sur la *syntagmatique* dans l'approche systémique, essentiellement parce que la notion de choix y est importante : par exemple le fait de choisir un type d'instance par rapport à un autre est considéré comme porteur de sens à part entière. Soulignons, de plus, que l'aspect syntagmatique permet de hiérarchiser les constituants dans une strate donnée afin d'expliciter les différentes fonctions actualisées par les instances choisies. Autrement dit, il facilite par exemple la schématisation de la construction grammaticale en termes de combinaisons syntaxiques. Il est donc important non pas pour la partie systémique de la théorie (en tant que système), mais plutôt pour l'aspect à la fois fonctionnel et compositionnel (vis-à-vis de la structure globale de la langue).

3.2.4 L'ordre paradigmatique (système)

Comme nous l'avons indiqué ci-dessus, aborder la paradigmatique repose sur l'appréhension préalable de la syntagmatique. Il s'ensuit de ce constat que certains semblent baser l'ensemble des propriétés intrinsèques de la paradigmatique sur l'axe dit vertical, par rapport à l'axe communément représenté comme linéaire ou horizontal de la chaîne parlée. Cette simplification schématise de manière prématurée ces deux termes en grammaire traditionnelle et prêterait rapidement à confusion si l'on transposait cette allusion à la conception systémique fonctionnelle de la langue.

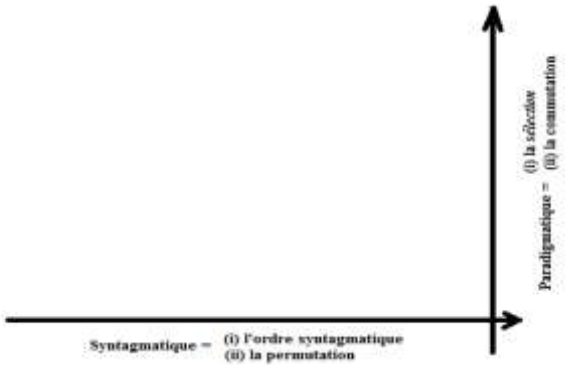


Figure 8: Les deux axes en grammaire traditionnelle

En grammaire traditionnelle donc, tout élément ayant un rapport associatif au sens saussurien avec un item sur l'axe linéaire est considéré comme appartenant à l'axe vertical. Ceci peut être explicité en reprenant l'exemple de l'énoncé de notre corpus introduit ci-dessus : à savoir *The cartoon*

[46] Une d'esquisse d'un cadre d'analyse syntagmatique (dit de *constituency*) est présentée de manière détaillée pour plusieurs strates dans Halliday & Matthiessen (2004) et Halliday (1994).

underlines environmental issues in our globalised economy. Le premier axe voudrait que l'on représente l'énoncé de la manière suivante en précisant les différentes catégories ou classes grammaticales: *Déterminant + Nom + Verbe + Adjectif + Nom + Préposition + Adjectif + nom*. Le deuxième axe, en prenant comme point d'ancrage le premier, suppose que l'on puisse librement commuter le déterminant *the* par un autre item appartenant à la même catégorie grammaticale. Il serait donc possible de procéder à une commutation avec *a, all, my, that, their, this,* et ainsi de suite. L'approche paradigmatique est perçue corollairement donc comme étant dépendante et liée aux items lexicaux figurant sur la chaîne linéaire.

Un premier point de démarcation de cette conception en linguistique systémique serait le fait que la paradigmatique ne constitue pas une approche singulièrement lexicale. C'est-à-dire que cet axe dépasse la simple commutation d'un mot avec un autre, voire d'un trait distinctif avec un autre, par exemple en phonétique. La notion d'*in absentia* que Saussure lui avait conférée se voit alors reprise et développée davantage chez les systémistes. En effet, dans la perspective systémique le système prime sur la structure. Le choix ou la sélection joue, de ce fait, un rôle primordial dans l'analyse linguistique – non seulement par rapport à ce qui a effectivement été choisi comme instance finale, mais également par rapport à l'ordre paradigmatique qui suppose ici que l'on s'intéresse à tous les choix qui ont été opérés, à ceux qui étaient possibles et ceux qui sont mutuellement exclusifs ou inclusifs.

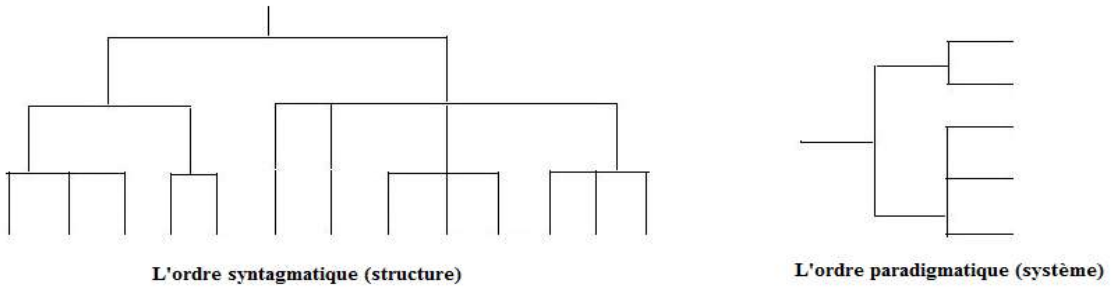


Figure 9 : Une conceptualisation en LSF montrant les deux axes en opposition

La figure 9 illustre la conception systémique des deux axes. D'emblée, l'on comprend pourquoi l'utilisation des termes d'axe vertical ou horizontal/linéaire n'est plus à même de schématiser l'étendue du phénomène à l'étude. Tout particulièrement, pour ce qui est de l'ordre paradigmatique, l'analyse de la sélection peut désormais se porter sur tout élément choisi comme instance finale : qu'il soit question d'un mot, d'un syntagme, d'une proposition ou de façon plus systémique des différentes possibilités offertes par les différentes strates notamment lexico-grammaticale et sémantique. La perspective verticale de l'ordre paradigmatique renvoie à l'ensemble des choix relativement associatifs à l'instance choisie, alors que le sens horizontal permet de signaler ce que

l'on appelle la *délicatesse* allant du plus général au plus spécifique (c'est-à-dire la dernière instance ou item choisi).

Pour expliquer davantage cette conception de sélection en LSF, il faut se rappeler que la sélection à laquelle on s'intéresse doit être étudiée en contexte : à savoir la situation contextuelle d'où l'item est issu. Et c'est justement ici que la notion de paradigmatique entre en jeu. En effet, le simple fait de choisir un mot par rapport à un autre ou encore un type de phrase (déclarative par opposition à interrogative) permet de bien construire un message – puisque le sens ne découle pas strictement de ce qui a été formulé, mais du choix que le système entier propose. Reprenons l'exemple de la phrase au sens traditionnel. Selon le sens que l'on veut donner à son message, la langue fournit deux options au départ, soit une phrase de type impérative soit indicative. Si l'indicatif est choisi, une nouvelle sélection s'impose, étant donné que chacune présente des spécificités propres. Enfin si l'interrogatif est retenu, la langue anglaise propose une dernière sélection entre des questions de type dites *yes* (y) ou *no* (n), ou encore des questions en *WH*. Ceci peut est schématisé de la façon suivante :

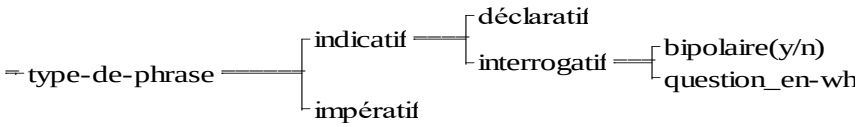


Figure 10 : Schématisation de la phrase interrogative en anglais

Toutefois les options de la langue ne sont pas toutes aussi simples. Examinons maintenant la figure 11. Pour que le type A se réalise sémantiquement, le locuteur (ou scripteur) choisit entre trois conditions d'entrée (*a1-a3*) ; par exemple si *a3* est choisi, la suite se porte sur ce qui est désigné comme étant première ou deuxième option – tout autre sélection entraînera une erreur de sélection. De même, *a2* et *a1* seront d'emblée exclus. Il serait de même si la condition *a2* avait été retenue, a1 et a3 seraient exclus.

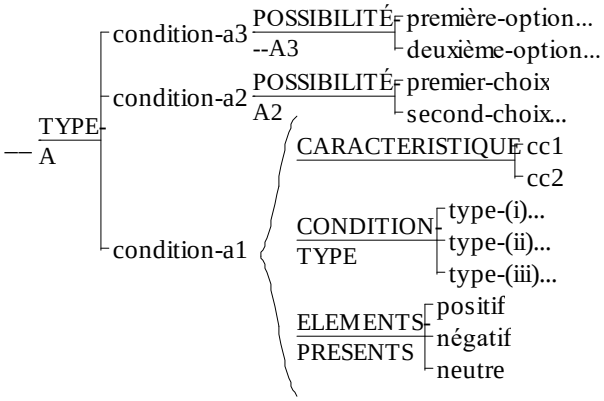


Figure 11 : Cheminement hypothétique des différentes conditions d'entrée de la langue

Les premiers types de sélection pourraient se formuler ainsi :

i. Type A => Condition *a3* => première option
ii. Type A => Condition *a3* => deuxième option
iii. Type A=> Condition *a2* => premier choix
iv. Type A => Condition *a2* => second choix

Toutefois, la condition *a1* n'opère pas de la même manière. Pour que celle-ci se réalise sémantiquement, la sélection n'est pas si directe ou évidente. Les trois sous-catégorisations qui constituent ensemble une seule et même « condition d'entrée » sont mutuellement inclusives et nécessaires pour que *Type A* – condition *a1* se réalise pleinement. Il faut donc choisir « un de chaque » pour que le tout crée du sens. A titre illustratif, un exemple pourrait être la grammaticalité et/ou l'acceptabilité de la valeur épistémique d'un énoncé en anglais avec « if ». Supposons donc que *Type A* renvoie à une « *it-clause* », condition a*1* pourrait être l'emploi du temps verbal, à la suite duquel on a le choix entre les différents types de conditionnel. Par exemple :

i. Type A => a*1* (if + présent simple = conditionnel zéro)
ii. Type A => a*1* (if + présent simple = 1er conditionnel *will*)
iii. Type A => a*1* (if + passé simple = 2ème conditionnel *would*)
iv. Type A => a*1* (if + plus-que-parfait = 3ème conditionnel *would have*)

Une illustration concrète de ces deux cas de figure, notamment exemples (i) et (ii) serait :

1) Even if diplomas help to get a job, there are still a lot of skilled workers [...] (txt_121_sm2)
2) If we do not find a solution, we will all share the environmental burden. (txt_010_sm1)

En bref, il convient de rappeler ici que l'aspect paradigmatique avec sa création visuelle de réseau ou nœud de réseaux est le point de départ de la théorie systémique et également son appellation. Ainsi, nous pouvons soutenir que cet aspect renvoie non pas à la verticalité des choix lexicaux pouvant permuter les uns avec les autres, mais à l'ensemble des choix qu'offre le système langagier dans les quatre strates définies. De ce fait, nous rejoignons donc Fontaine (2013) qui considère la notion de choix en LSF comme fondamentalement axiomatique au point où cette notion n'est que très peu remise en question au vu de son statut d'*acquis*.

3.2.5 Les métafonctions

Comme indiqué ci-dessus, nous rejoignons la position de Matthiessen (2007) qui soutient que le cadre systémique doit être entendu comme étant évolutionnaire plutôt qu'une théorie révolutionnaire. Et ce, notamment en raison du fait qu'il est fondé sur des concepts considérés comme embryonnaires et pré-modélisés. Ces concepts développés séparément se sont avéré

foncièrement complémentaires, et c'est cette complémentarité qui a donné naissance au cadre holistique que l'on connait aujourd'hui. Cependant, tout n'est pas que le fruit d'une évolution constante : ce qui est nouveau et unique à la LSF – c'est-à-dire qui puise son origine dans le modèle systémique - est ce que l'on appelle les métafonctions. Ces métafonctions constituent donc le cinquième et dernier élément du modèle architectural de la langue que nous aborderons dans cette sous-section.

Tout d'abord, le terme *métafonction* renvoie à des ressources sémantiques intrinsèques à la langue, à la grande différence des fonctions dites grammaticales. En effet, ce terme désigne la valeur représentative qui est véhiculée à travers chaque emploi spécifique d'un énoncé ou dans la terminologie proprement LSF, une proposition. De ce fait, il facilite le repérage des différentes valeurs de représentation attribuées aux nombreux éléments de la proposition, nous permettant par la suite d'identifier les degrés subtils de précision codés de manière inconsciente et simultanée par chaque locuteur ou scripteur. Notons que Halliday & Matthiessen justifient l'emploi du terme systémique de *métafonction* par opposition au terme très usité en grammaire de *fonction* de la manière suivante :

> [H]owever, there is a long tradition of talking about the functions of language in contexts where 'function' simply means purpose or way of using language, and has no significance for the analysis of language itself (cf. Halliday and Hasan, 1985: Chapter 1; Martin, 1990). But the systemic analysis shows that functionality is **intrinsic** to language: that is to say, the entire architecture of language is arranged along functional lines. Language is as it is because of the functions in which it has evolved in the human species. The term'metafunction' was adopted to suggest that function was an integral component within the overall theory. (2004 : 30-31)

Ces derniers en sont arrivé à cette terminologie en voulant préciser davantage les différentes fonctions du langage – notamment en se démarquant de leurs prédécesseurs (Halliday & Hasan, 1985). En effet, dans un chapitre sciemment intitulé « Functions of Language », ces derniers repassent en revue les travaux de Bühler, Jacobson, Malinoski, pour ne citer que quelques-uns, ayant tous abordé de manière différente la « fonction du langage ». Halliday & Hasan rappellent ici que pour de nombreux auteurs ce terme était entendu uniquement en termes d'usage effectif d'une unité linguistique, ce qui impose d'emblée une certaine limite dans l'appréhension linguistique globale de ce concept. C'est ainsi que Halliday & Hasan postulent que dans une approche qui se veut tout d'abord fonctionnelle, ce concept doit pouvoir dépasser ce cadre d'usage effectif pour s'étendre à l'ensemble du système linguistique – en intégrant notamment une dimension sémantique.

> Function will be interpreted not just as the use of language but as a fundamental property of language itself, something that is basic to the evolution of the semantic system (1985 : 17)
>
> There has been a lot of misunderstanding of the concept of functions of language. It has often been assumed that each sentence has just one, or at least one primary, function; or, even if the sentence is recognised to be multifunctional, that it ought to be possible to point to each separate part of the sentence and to say this part has this function, and that part has that function, and the other part has the other function [...] language is certainly not like that. Every sentence in a text is multifunctional; but not in such a way that you can point to one particular constituent or segment and say this segment has just this function. The meanings are woven together in a very dense fabric ... (ibid.: 23)

Cela étant, Bloor & Bloor fournissent une entrée définitoire succincte de ce terme que nous reprenons ci-dessous. Leur définition qui a le mérite d'être claire conçoit les métafonctions comme « *[o]ne of the three superordinate functional categories which characterize meaning in language. These are: ideational, interpersonal and textual. They co-exist in all texts* » (2004 : 284). Ici le principe de "superordinate functional categories" puise sa raison d'être au niveau sémantique et non dans une catégorisation grammaticale. Les trois métafonctions peuvent donc se résumer de la façon suivante :

- La métafonction idéationnelle permet de parler de la réalité expérientielle principalement (i) en nommant les entités, les événements et les états dans le discours (les participants) ; (ii) en reliant les participants entre eux ou des propositions entre elles (fonction logique) (iii) en repérant ceux qui ne peuvent qu'indirectement occuper la fonction de participant (les circonstances) ; (iv) et en mettant le tout en rapport avec le noyau verbal (le procès).
- La métafonction interpersonnelle reconstitue les rapports sociaux entre locuteurs et allocutaires. Il permet par exemple de comprendre la neutralité ou la polarité d'un message ainsi que la posture du locuteur tant par rapport à son message qu'à son allocutaire.
- La métafonction textuelle assure une dimension constructiviste dans la mesure où elle facilite l'organisation discursive en garantissant singulièrement la cohérence et la cohésion de l'ensemble.

Notons toutefois qu'à travers une analyse des trois métafonctions, il est possible d'une part de s'intéresser au sens final réalisé par chaque métafonction, et d'autre part d'examiner les ressources proprement linguistiques mobilisées dans les différentes métafonctions. Le premier cas de figure se réalise dans ce que l'on pourrait comparer à l'analyse du discours, puisque le focus n'est véritablement pas les moyens employés pour créer le message, mais plutôt le sens véhiculé. Etant

donné que nous ne nous inscrivons pas dans une telle approche, notre présentation et, par conséquent, notre application des métafonctions se limitera donc aux seules ressources linguistiques nécessaires pour l'analyse des différentes métafonctions.

La métafonction idéationnelle

Un cadre théorique du langage qui se veut plurivalent - en s'inscrivant dans une approche à la fois sémiotique, sociale et fonctionnelle - suppose avant tout que l'usage premier de la langue réside dans l'idéation. C'est-à-dire le fait de créer, et singulièrement ici, de communiquer des idées. Derrière le principe d'idéation, il s'ensuit que ces idées sont formulées en propos plein de sens, en termes de contenu et d'intelligibilité pour le récepteur/interlocuteur. Il importe, de ce fait, que les usagers d'une langue aient recours aux mêmes moyens, non seulement en termes de lexique mais plutôt en ce qui concerne le « comment » représenter et partager une information. Ces réflexions dans notre cadre théorique renvoient à la métafonction idéationnelle qui, de manière générale, constitue une fonction essentielle de la langue. En effet, Halliday & Matthiessen (1999) estiment que cette métafonction est de la première importance dans la mesure où c'est à travers celle-ci que l'on construit et appréhende la réalité. Ces derniers vont jusqu'à soutenir que « grammar is a theory of human experience; it is our interpretation of all that goes on around us, and also inside ourselves ». De même, Eggins (2004) nous donne une esquisse de définition très claire en affirmant : « ideational meanings [is] about how we represent reality in language. »

Malgré l'abstraction épistémologique apparente derrière ces affirmations, nous reconnaissons que la métafonction permet une schématisation utile de la langue. Selon les différents moyens linguistiques mis en œuvre par un locuteur, il est question ici d'extraire une représentation fidèle de l'idéation à travers le rapport qu'entretient le locuteur tant avec le réel qu'avec la langue. Pour ce faire, Halliday distingue deux facettes dans cette métafonction : la première dite *expérientielle* et la deuxième *logique*. Notons cependant que nous détaillons la première facette de manière approfondie sans trop nous attarder sur la deuxième ; et ce, étant donné que cette dernière n'est que très peu exploitée dans notre cadre méthodologique (cf. chapitre IV) et conséquemment dans le cadre analytique. Il convient néanmoins de dire que celle-ci occupe une fonction de « liage » au niveau des propositions complexes et qu'elle sous-tend les structures dites hypotactiques et paratactiques (cf. chapitre VIII).

Il est à souligner que la métafonction expérientielle constitue l'un des concepts phares de la théorie générale des métafonctions. Cette importance devient rapidement manifeste dès lors que l'on examine sa structure analytique qui s'est développée sous forme d'un système de réseaux, qui, de

plus, présente un large éventail de choix subordonnés les uns aux autres. En substance, cette métafonction est analysée à travers ce que l'on nomme la transitivité. Ceci n'est pas à confondre avec le fait qu'un verbe soit transitif ou intransitif en grammaire traditionnelle. En effet, à la base de ce niveau d'analyse l'on notera les trois constituants principaux qui sont explicités par le biais du principe de l'ordre paradigmatique (ce principe est expliqué à la section 3.2.4) Ces trois premiers composants sont : (i) le procès, (ii) le(s) participant(s), et (iii) la/les circonstance(s). Dans cette perspective expérientielle, le premier élément nous renseigne sur la nature de l'événement en cours ; le deuxième sur la nature ou le rôle des entités impliquées dans l'événement ; et le troisième apporte de manière indirecte des précisions supplémentaires, tout particulièrement en qualifiant ou spécifiant davantage le procès. Nous devons reconnaître toutefois que ce niveau d'analyse existe dans bien des théories en grammaire traditionnelle, où les deux premiers éléments sont étiquetés (i) Sujet/Objet/Acteur et (ii) verbe.

Ce qui pousse notre cadre actuel à dépasser cette classification est tout simplement le fait que la grammaire traditionnelle nous « prive » d'informations de types très variés portant non seulement sur la langue en tant que potentiel linguistique, mais en tant que système actualisé et représentatif de ses usagers. De plus, ces informations sont hautement codées, mais n'en demeurent pas moins identifiables à travers une analyse appropriée. Ainsi, prenons l'exemple du fait d'étiqueter un élément « *verbe* » et/ou « *auxiliaire* », sans davantage de catégorisation possible – en fonction de leur différence notamment sémantique et typologique. Dans notre cadre actuel, le terme *verbe* est remplacé ou plutôt est encodé par *procès* et celui-ci se décline en six types différents – nous apportant des précisions fort utiles pour l'analyse qui va suivre. Ces distinctions ont des incidences directement sur les composants de la proposition ; en ce qu'ils se traduisent par des types précis de participants par rapport au procès. Il en va de même pour les plusieurs types de circonstances.

Examinons maintenant la figure 12 ci-dessous dans laquelle l'on identifie les trois composants principaux de la transitivité. Rappelons que ces éléments peuvent s'actualiser de manière simultanée dans une même proposition, sans aucune obligation de les avoir tous les trois en même temps. Le procès en LSF se décline en six types différents : à savoir, matériel, comportemental, mental, verbal, relationnel et existentiel.

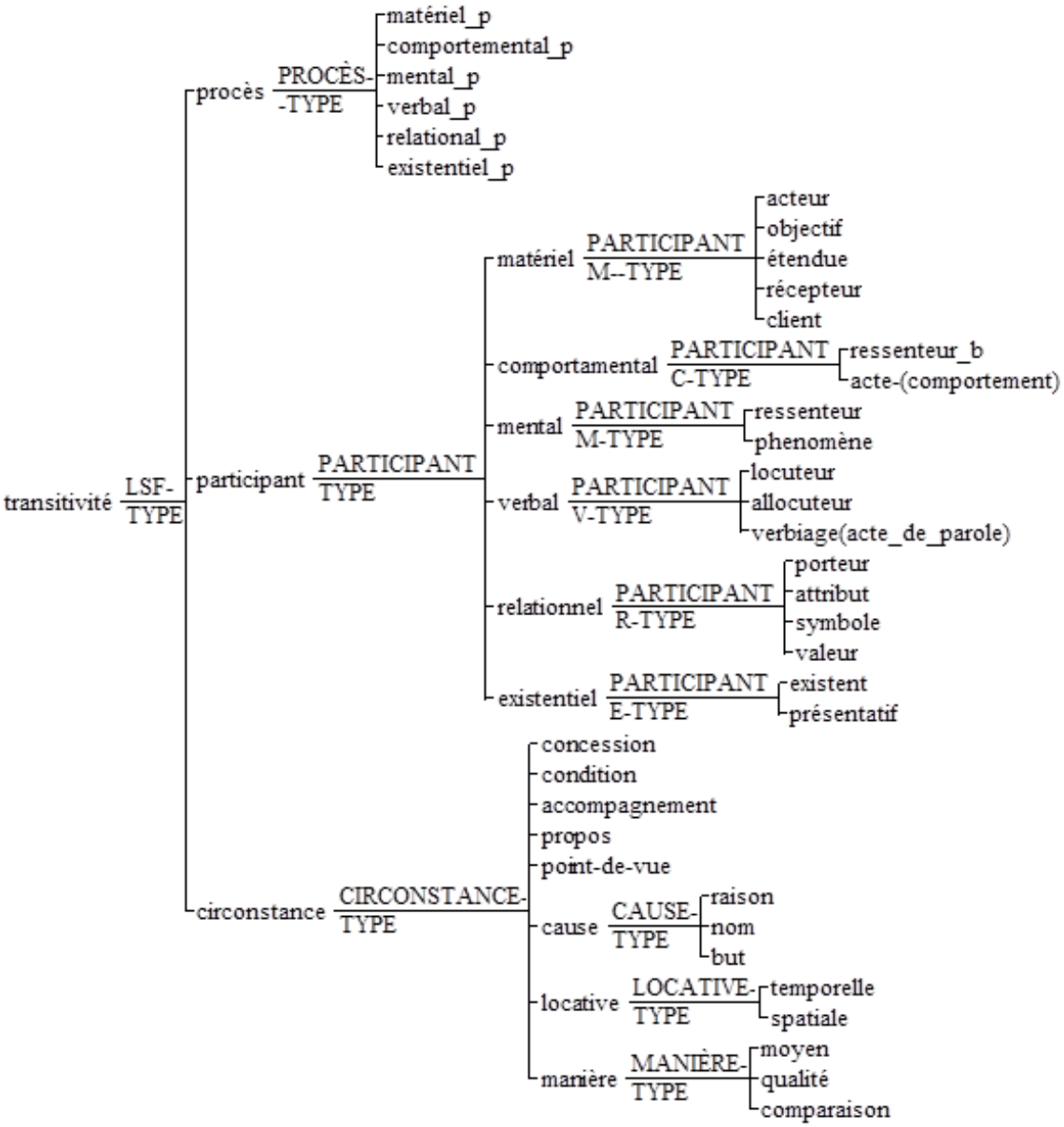


Figure 12 : Schéma simplifié des constituants de la Transitivité

Nous reprenons les très brèves descriptions faites par Banks (2005) pour les six procès :

i. Les procès matériels […] expriment des changements dans le monde physique. Ainsi, « marcher », « lancer », « donner », sont normalement des procès matériels.

ii. Les procès comportementaux se trouve[nt] à la frontière des procès matériels et des procès mentaux. Ils comportent souvent des traits psychologiques ou physiologiques, parfois les deux combinés. Ainsi, « sourire », « renifler », « ronfler », seront typiquement des procès comportementaux[47].

[47] Ce terme a été traduit en français par « réactionnel » (Banks, 2005), mais le terme « comportemental » nous semble plus adéquat de façon à limiter les interprétations trop larges tout en restant fidèles au sens d'origine.

iii. Les procès mentaux concernent les événements qui ont lieu au niveau cérébral. On peut y distinguer trois groupes : les procès cognitifs, comme « penser », « croire », « considérer », les procès de perception, comme « voir », « entendre » ; et les procès affectifs, comme « aimer », « adorer », « détester ».
iv. Les procès verbaux sont des procès de communication, que ce soit la communication orale ou la communication écrite.
v. Les procès relationnels expriment les relations entre deux entités, ou entre une entité et une propriété, sans événement ou changement physique… [Ils] expriment des états ; ainsi, « être » et « avoir » sont des procès relationnels prototypiques. On inclut ainsi dans la même catégorie les procès de devenir, procès qui expriment la création d'un état.
vi. Les procès existentiels expriment l'existence d'une entité.

Etant donné que le choix d'un type de procès spécifique restreint le choix des participants avec lequel il est compatible, nous dressons ci-après à titre indicatif[48] une liste des participants qui fonctionnent comme des satellites – et ce, indépendamment de l'ordre de combinaison. Dans ce cas, supposons que pour chaque procès nous avons affaire à un ou plusieurs participants : un qui permet au procès de s'actualiser et les autres qui se voient affectés ou modifiés par le procès ou apportent des spécifications sur ce dernier. Soulignons également ici que le rôle des participants n'est pas figé et mais interchangeable : par ailleurs une proposition peut comporter un, deux ou plus de ces candidats s'enchaînant les uns après les autres, notamment dans la rubrique intitulée « Participant 2 ».

(i) Procès matériel :

Participant 1	Procès	Participant 2
acteur	pro : matériel	objectif
acteur	pro : matériel	étendue
acteur	pro : matériel	récepteur
acteur	pro : matériel	client

(ii) Procès comportemental :

Participant 1	Procès	Participant 2
ressenteur	pro : comportemental	acte

(iii) Procès mental :

Participant 1	Procès	Participant 2
ressenteur	pro : mental	phénomène

(iv) Procès verbal :

Participant 1	Procès	Participant 2

[48] Cf. Banks (2005) et Cafferel-Cayron (2009) pour une liste exhaustive en français des participants satellites.

locuteur	pro : verbal	allocutaire
locuteur	pro : verbal	verbiage

(v) Procès relationnel :

Participant 1	Procès	Participant 2
porteur	pro : relationnel	attribut
symbole	pro : relationnel	valeur

(vi) Procès existentiel :

Participant 1	Procès
Existentiel	pro : existentiel
Présentatif	pro : existentiel

Quant au troisième et dernier composant de la strate expérientielle, à savoir les circonstances, il nous parait judicieux de rappeler ici que ces dernières sont en quelque sorte des électrons libres en ce qu'elles peuvent en grande partie se combiner librement avec tous les participants et tous les procès. Cela ne veut pas dire pour autant qu'il n'y a pas de contraintes régissant leur emploi : mais c'est justement cet effet de « liberté » qui semble constituer un piège subtil pour un grand nombre d'apprenants, comme nous le verrons dans les chapitres V et VI.

La métafonction interpersonnelle

Si l'on admet que la langue fournit les moyens d'appréhender le réel, en nous permettant de nommer différentes entités afin de les classifier et de les mettre en rapport les unes avec les autres – tout particulièrement autour d'un noyau verbal appelé ici *le procès* – elle nous permet également de traduire et représenter les relations sociales qu'entretient un locuteur précis avec la société et son allocutaire. Ces représentations se réalisent au niveau de la métafonction interpersonnelle. Mais comme pour la métafonction expérientielle, les ressources pour actualiser ces fonctions résident dans les ressources proprement lexicogrammaticales de la langue. Toutefois notons d'emblée que la métafonction interpersonnelle dépasse les éléments que nous développerons ci-dessous. En effet, selon Halliday & Matthiessen (1999, 2004) le but principal de cette métafonction est de représenter à la fois les relations sociales et le message (représenté ici par la proposition) comme moyen d'échange au sens de partage d'information. A titre d'information donc, signalons simplement qu'au niveau discursif le cadre interpersonnel s'intéresse également aux actes de langage selon qu'il est question de proposer, commander, affirmer, questionner et ainsi de suite.

Mais le cadre qui nous intéresse tout particulièrement ici est la schématisation grammaticale – puisqu'il s'apparente d'une manière générale à la syntaxe en grammaire traditionnelle tout en permettant d'approfondir les différentes fonctions syntaxiques existantes. Pour les besoins donc de

notre analyse, disons que la métafonction interpersonnelle à travers son système de mode (cf. la figure 13) facilite la description linguistique à quatre niveaux qui peuvent s'actualiser dans une même proposition de manière simultanée.

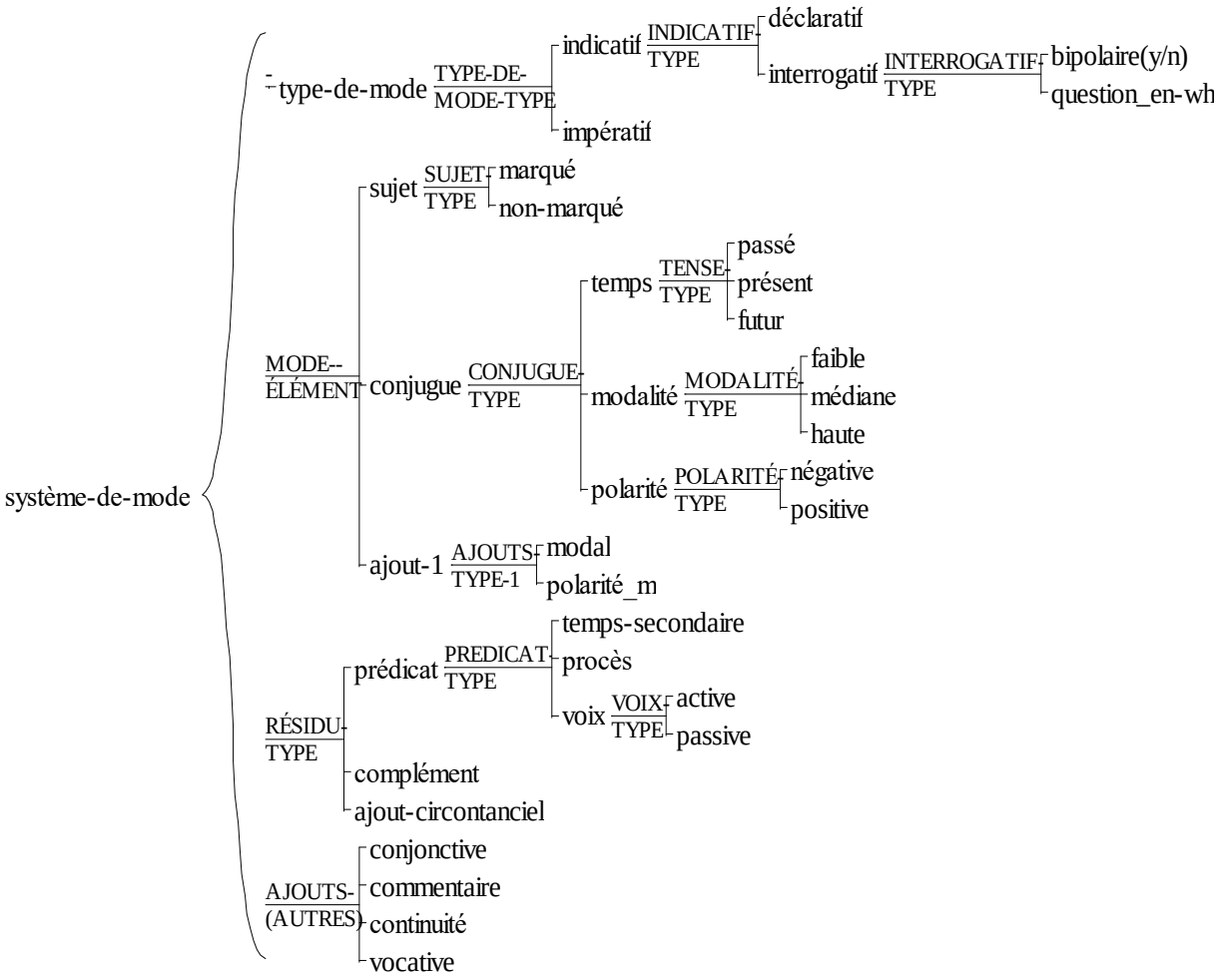


Figure 13 : Le système de mode en LSF

i. Tout d'abord, il s'agit ici de différencier les types de proposition que l'on examine : c'est-à-dire, selon que la proposition sous étude est indicative ou impérative ; si elle est indicative le choix se pose entre déclarative et interrogative ; et si elle est interrogative, le choix se pose entre des questions en *yes/no* (ou pour reprendre la terminologie de Banks (2005) de type *bipolaire*), soit en *WH*.

ii. Le système de mode se divise en deux parties : Mode et Résidu. Le Mode se subdivise en *sujet* et *conjugue*[49]. Le sens traditionnel de sujet demeure quasi-inchangé, mais celui de conjugue renvoie à « la partie 'conjuguée' du verbe […] qui attire les marques de temps et de nombre » et la négation (Banks 2005).

[49] La traduction des certains termes –d'anglais au français – a été réalisé dans les travaux de Banks (2005, 2009).

iii. Le Résidu est à mettre en opposition par rapport au Mode. Il renvoie à ce qui n'est pas compris dans le Mode – à savoir le prédicat appelé aussi le verbe lexical ou principal y compris le complément et certains types d'ajout, notamment circonstanciel.
iv. Le quatrième élément relève des ajouts de types divers.

La figure 13 fournit donc un aperçu simplifié du système de la métafonction interpersonnelle telle qu'elle est conçue en linguistique systémique fonctionnelle. Et ce modèle réduit constitue celui que nous appliquons à notre analyse. Etant donné que les quatre nœuds principaux - à savoir (i) le type de modes, (ii) les éléments qui composent le Mode, (iii) ceux qui sont présents dans le Résidu (iv) et les ajouts - peuvent s'actualiser simultanément, il est important cependant de se rappeler ici qu'il n'y a pas de hiérarchisation entre ces éléments.

Lors que l'on parle de Mode (ou *Mood* en anglais dans la littérature systémique), en faisant allusion à l'ensemble des ressources disponibles dans la métafonction interpersonnelle, les quatre nœuds ne sont pas traités de la même manière : il est souvent uniquement question de *Mode* et de *Résidu*. Les premier et quatrième éléments sont donc fréquemment exclus de l'équation. Et pour cause : ils ne sont pas toujours présents ou jugés importants. Par exemple dans la proposition ci-dessous le choix du déclaratif peut ne pas être primordial dans l'analyse, ou encore le fait qu'il n'y a pas d'ajout que nous appellerons pour le besoin de brièveté du *type-2* – c'est-à-dire des ajouts qui se réalisent sémantiquement à l'extérieur du Mode et du Résidu. Ceci amène donc certains à traiter ces deux composants comme les seules propriétés de la métafonction interpersonnelle.

Interpersonnel (MODE)	Mode			Résidu					
	Sujet		*Conjugue /prédicat*	complément		Ajout			
	The	cartoon	underlines	environmental	issues	in	our	globalised	economy

Figure 14 : Exemple d'analyse interpersonnelle

En outre, nous pouvons de manière succincte affirmer que le Mode – en tant que partie -comprend donc trois éléments potentiels : les deux premiers sont obligatoires, à savoir le sujet et le conjugue. Le sujet est celui identifiable principalement par la reprise d'une *question tag* ou le *tag test* (cf. Fawcett 2005) - du type *You are from Paris, aren't you ?* Le conjugue renvoie donc à l'élément apportant des précisons sur le nombre, le temps, la modalité et la polarité : de ce fait il est souvent représenté par l'auxiliaire dès lors qu'il s'agit d'un temps composé en anglais. Toutefois notons que le conjugue peut être amalgamé avec le prédicat dans les formes telles que les passé et présent simples.

1	The vase is broken, isn't it?	le conjugue est séparé du prédicat
2	He is home, isn't he?	le conjugue et le prédicat sont amalgamés
3	He should have called, shouldn't he?	le conjugue est séparé du prédicat

Tableau 10 : Exemple de séparation entre conjugue et prédicat

Dans ces trois exemples la partie dite de « question tag » renvoie dans le premier exemple au conjugue *is* qui est séparé du prédicat. Nous pouvons relever aussi le fait que la négation et le sujet *The vase* repris par *it* sont également présents dans cette structure – et ce, sans le verbe dit lexical ou principal. Cela laisse entendre que ce dernier occupe une tout autre fonction dans la phrase. Nous pouvons procéder à une analyse similaire avec les deux autres énoncés ; soulignons que dans l'énoncé n°3 l'auxiliaire *should* se voit repris dans la question tag, ce qui nous permet de l'assimiler tout en le distinguant de la fonction conjugue. Notons ici, pour des besoins de précision, que le conjugue accepte en son sein les auxiliaires modaux et les adverbes de modalité allant d'une modalité faible (cf. *might*) à une modalité haute (cf. *must, certainly*) Eggins (2004).

Quant aux ajouts « autorisés » dans la partie « mode – élément » de la métafonction, Eggins (2004) nous affirme qu'ils sont de deux types : (i) modal et (ii) polarité : pour des besoins de brièveté, disons également que ces deux éléments sont quasi-interchangeables avec les auxiliaires dits modaux et les éléments qui forment la négation en anglais (à savoir *will, would, shall* ; et soit *do not* le cas échéant, soit *not* employé seul).

Dans la partie dit Résidu, il est possible d'identifier tout d'abord le procès qui nous apporte plusieurs précisions. La première porte sur ce que nous appelons ici le temps secondaire. Dans l'exemple *The name of the royal baby has been officially announced* "has" occupe la fonction de conjugue et "been" celle de temps secondaire, tandis que « announced » occupe la fonction de procès, appelé verbe lexical en grammaire traditionnelle. Le temps secondaire a donc le rôle d'apporter des précisions à valeur aspectuelle, comme dans notre exemple. Ajoutons ici qu'il est également possible de repérer la « voix » employée à travers l'identification du procès : à savoir entre active et passive. Notons enfin que dans le Résidu, l'on retrouve également les compléments et certains types d'ajouts. En effet, comme dans la partie « mode élément », le Résidu accepte certains types d'ajouts en son sein : à savoir ici uniquement ceux de types circonstanciels.

La métafonction textuelle

La troisième et dernière métafonction, comme son nom l'indique, fonctionne non seulement à un niveau lexical ou simplement propositionnel, mais trouve toute sa pertinence au niveau proprement textuel. Cela signifie que celle-ci s'articule essentiellement autour de la structure discursive ou textuelle dans laquelle cette métafonction assure une fonction constructiviste ; tout particulièrement

dans la mesure où elle se veut garante de la cohérence et de la cohésion de l'ensemble. En effet, comme le soulignent Halliday & Matthiessen aucune des deux constructions dite idéationnelle ou interpersonnelle n'aurait de sens sans cette métafonction.

> The "textual" metafunction is the name we give to the systematic resources a language must have for creating discourse: for ensuring that each instance of text makes contact with its environment. The "environment" includes both the context of situation and other instances of text. Relative to the other metafunctions, therefore, the textual metafunction appears in an enabling role; without its resources, neither ideational nor interpersonal constructs would make sense (1999: 528).

Cela étant, la fonction inhérente ici est de tisser le texte : c'est-à-dire de faire en sorte d'établir les relations de dépendances ou de liage nécessaires pour la compréhension globale des différents éléments qui se succèdent dans ledit texte. Il importe donc de labeliser ces éléments selon la valeur à la fois thématique et informationnelle qu'ils apportent à la structure propositionelle : nous identifions donc deux composants de base, à savoir le thème et le rhème. Le thème ici n'est entendu ni dans le sens de Lambrecht (1994), ni dans celui privilégié par Mathesius[50] du cercle linguistique de Prague, mais plutôt dans celui avancé par Halliday. Et pour cause, nous ne négligeons pas les apports théoriques des deux premiers[51], mais notons simplement que leurs spécificités ne sont pas intégralement compatibles avec notre cadre d'analyse et, de plus, ne sont pas porteurs en eux-mêmes de catégorisations jugées utiles et exploitables dans la présente étude.

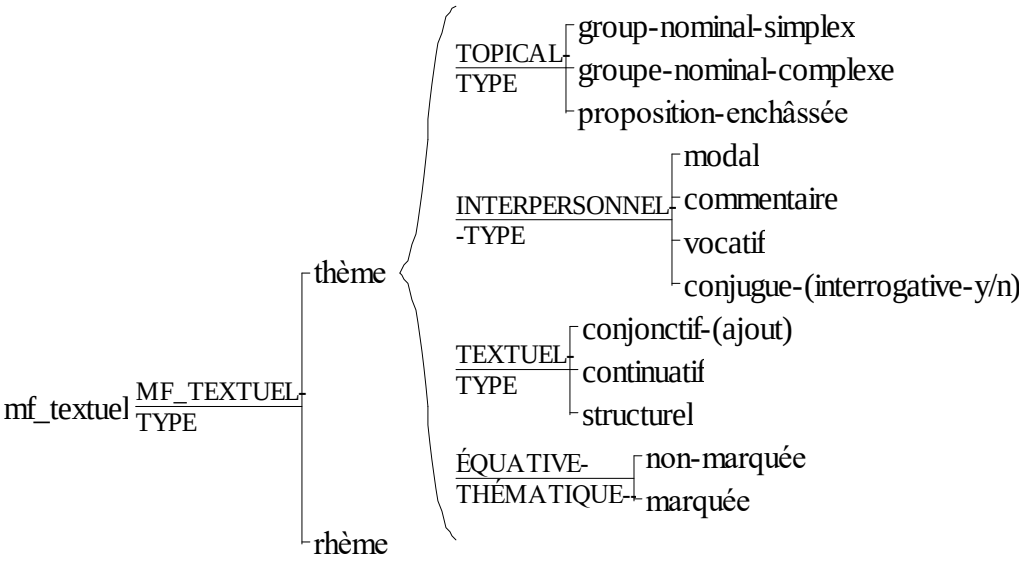


Figure 15 : le système de métafonction textuelle

[50] Cf. Nekula (1999) au sujet des travaux de V. Mathesius, co-fondateur du cercle linguistique de Prague.
[51] Cf. Carter-Thomas (2009b) pour une discussion sur les apports de ces deux cadrages théoriques.

Le thème correspond donc ici aux éléments situés à l'extrême gauche de la phrase ou au point de départ dans la proposition. Par opposition le rhème constitue le reste ou ce qui est dit à propos du thème. Cette définition simple sans être simpliste nous permet de schématiser l'ensemble de la métafonction textuelle de la manière suivante :

- Le thème topical renvoie à un des trois éléments de base de la métafonction expérientielle qui constitue de ce fait un élément obligatoire (Halliday & Matthiessen 2004, Banks 2005)
- Le thème interpersonnel, lui, n'est pas obligatoire. Les ajouts dits modal, commentaire et vocatif sont représentés ici. Le conjugue en position initiale de la proposition fait également partie de ce thème.
- Le thème textuel est selon Banks (2005) un élément thématique qui crée un lien avec le co-texte. Il est donc question ici « des marques de conjonction » et « [l]es éléments de liaison discursive ».
- Les équatives thématiques sont plus tranchées, par rapport aux systèmes très élaborés des différents thèmes. En effet, elles sont sans éléments hiérarchisés et ne renvoient qu'à un type de proposition donnée. Ladite proposition s'apparente à une structure quasi-mathématique dans laquelle le thème est égal au rhème (thème=rhème). Cette structure rappelle le procès relationnel, notamment à travers son principe de *valeur*. Précisons, à titre d'information, que les équatives thématiques sont aussi appelées des phrases pseudo-clivées ou wh-clefts et jouent un rôle très important sur le plan informationnel (cf. Carter-Thomas & Rowley-Jolivet 2001).

3.3 Quelques cas d'utilisations de la LSF réalisés avec corpus

Étant donné que la linguistique systémique fonctionnelle est devenue une théorie du langage que nous pouvons qualifier de circonstanciée tout en étant extensible, son cadre conceptuel se voit désormais appliqué à des domaines comparativement divers et variés ; et ce, dès lors que la langue en tant qu'élément central soulève une question ou pose un problème. De ce fait, il est à souligner que l'application de son cadre n'est nullement limitée au champ générique des sciences du langage, en tant que méthode d'analyse. En effet, notre cadre cultive désormais une approche multiforme dans la mesure où il s'emploie non seulement pour faire de la description linguistique, mais assure parallèlement un côté pratique – hors du milieu de la théorisation proprement linguistique. Cela a donné lieu à l'émergence d'une théorie dite applicable (*appliable*, dans la littérature anglo-saxonne ;

Webster 2013 ; Mahboob & Knight 2010), renvoyant ainsi à son utilité ou plus précisément son applicabilité à l'ensemble des aspects où la langue constitue au moins l'une des problématiques principales à l'étude : allant donc de l'analyse grammaticale aux études dites littéraires ou discursives (comprendre, l'analyse du discours, l'analyse ou critique littéraire et ainsi de suite), sans oublier les études où l'homme fait partie intégrante de l'objet d'étude par exemple dans le cas de la neurolinguistique, la psycholinguistique, et notamment la sociolinguistique pour ne citer que quelques unes.

Dans cette optique, nous suggérons que le cadre LSF dépasse le clivage général[52] de la linguistique dite théorique d'un coté et appliquée de l'autre dans la tradition anglo-saxonne. Ceci est dû notamment au fait que le cadre LSF a été employé (i) pour des questions allant de la conceptualisation des universaux linguistiques et la modélisation de la langue comme système sémiotique et (ii) à celle de l'application de ces mêmes modélisations tant à des données qu'à des pratiques réelles et effectives dans de nombreuses langues. Par conséquent, ce constat nous amène à soutenir que la distinction dichotomique entre la linguistique théorique et la linguistique appliquée n'est pas un concept particulièrement transposable au modèle LSF, en raison du fait qu'aucun des deux n'est singulièrement privilégié. Nous pouvons cependant dresser une esquisse succincte du champ d'application de notre cadre théorique en faisant une distinction possible[53] entre les études réalisées avec ou sans corpus, ce qui relève de la théorisation et de la modélisation linguistique, ou encore ce qui relève des applications pratiques notamment dans le domaine de la pédagogie.

3.3.1 En didactique des langues maternelles et étrangères

Malgré le fait que nous réfutons la distinction réductrice entre linguistique théorique et linguistique appliquée, nous employons ici le terme générique de didactique des langues étant donné sa spécificité dans notre contexte français. En effet, dans la tradition anglosaxone[54] la distinction entre linguistique théorique et appliquée n'est pas transposable en l'état aux termes français de sciences du langage et la didactique des langues ; et ce, malgré l'apparente ressemblance conceptuelle. Une

[52] Cf. Beaugrande (1998) pour une discussion dans laquelle il critique l'opposition faite par H.G. Widdowson entre linguistique appliquée et linguistique théorique. En effet, ce dernier dresse une liste de composants rangés dans un des deux champs donnés ; toutefois, si l'on les étudie attentivement il devient alors possible d'affirmer que le cadre LSF n'est en aucun cas limité à un coté plus qu'un autre – mais s'avère clairement exploitable par l'ensemble des sous-domaines ou traits listés.

[53] Cette distinction n'est faite ici qu'à des fins de comparaison et ne constitue pas en soi une distinction courante dans le cadre proprement systémique.

[54] Nous reconnaissons que dans la littérature anglo-saxonne ces terminologies sont loin d'être homogènes, mais nous renvoyons à Bussman (1996) pour une définition succincte qui a le mérite d'être très englobante sans tomber dans des inexactitudes.

différence significative réside dans la délimitation de la didactique des langues (et donc, dans la tradition française) entendue comme une activité principalement pédagogique – dans un contexte d'enseignement-apprentissage donné. Or la linguistique appliquée (dans la tradition anglosaxone) renvoie non seulement à l'enseignement, mais à toute application pratique et interdisciplinaire employant des théories linguistiques : par exemple dans l'identification et le traitement des pathologies du langage ou encore pour les questions ayant trait à la traductologie, entre autres.

Par ailleurs, comme nous l'avons expliqué dans la section 3.1.1, l'origine de notre modèle théorique réside dans une série de rencontres causales entre des véritables besoins de descriptions linguistiques correlés avec des objectifs proprement didactiques, à savoir quand Halliday cherchait à enseigner le chinois en tant que langue étrangère. Ceci nous amène au constat que la didactique en tant que telle occupe toujours une place prépondérante dans la conception systémique – à la fois comme constituant un moteur derrière la description linguistique tout en étant une de ses finalités d'application. Cette dualité est particulièrement mise en avant quand de Beaugrande (1998) souligne à propos du cadre LSF que « it resolutely seeks to integrate theory with practice, and to place theoretical issues in the context of potential applications ».

Nul doute donc, comme nous avons cherché à le démontrer tout au long de ce chapitre, que le cadre systémique constitue une théorie qui s'est imposée à nous de par sa robustesse et singulièrement en raison du fait que son objectif principal réside dans la description complète de toute manifestation effective de la langue, quel que soit son contexte. Cela étant, le passage d'une théorie du langage abstraite à une théorie fonctionnelle, au sens large, demeure, à notre sens, dans l'application pratique de cette dernière à des données susceptibles non seulement de prouver ou réfuter son bien-fondé mais avant tout de démontrer son efficacité de manière concrète. Pour ces raisons nous nous intéressons aux utilisations pratiques du cadre LSF, loin de toute conceptualisation théorique.

De ce fait, nous portons une attention particulière à l'Australie[55]. En effet, étant donné que le cadre LSF a été développé en grande partie dans le milieu universitaire australien, le contexte éducatif local était propice pour vérifier quelques unes des hypothèses avancées. Ceci a conduit tout naturellement à une série de tâtonnements auprès des institutions éducatives du pays au cours des trente dernières années (cf. Christie 2004). Aujourd'hui, de nombreuses études font état de l'étendue de l'application pratique de LSF, non seulement dans les cours théoriques, mais dans la

[55] Il convient de signaler, que bien qu'ayant débuté au Royaume Uni dans les années 1960, le développement de la LSF s'est largement poursuivi en Australie avec la nomination de Halliday et d'autres collègues « systémistes » dans des universités australiennes.

formation des enseignants de langues et à son adaptation et exploitation en matière de pédagogie dans le milieu de l'enseignement primaire, secondaire et universitaire. Comme il s'agit d'une théorie du langage, nous notons toutefois que la zone la plus sensible de son utilisation pratique – dans le contexte australien – s'est illustrée dans l'amélioration générale de l'écriture, entendue à la fois au sens de la rédaction et dans les progrès en lecture.

En langue maternelle

Dans les années 1980, des spécialistes (linguistes et didacticiens) ont fait l'état des lieux de la situation locale du système éducatif australien – et le constat a été frappant. Il y avait une disparité non-négligeable chez les écoliers, selon leurs origines sociales (Rose 1999 ; Martin & Rose 2005 ; Acevedo & Rose 2007). Certains, notamment les aborigènes et ceux issus de milieux défavorisés étaient en quelque sorte délaissés par le système éducatif – et avaient un niveau scolaire inférieur à ce qui était attendu à leur âge. C'est à partir de ce constat que certains ont mené des réflexions sur l'amélioration des compétences en lecture et en écriture de ces jeunes australiens, jugés « à risque ». Le programme « *Learning to Read: Reading to Learn*[56] » est issu de cette initiative.

Ce programme prend comme point de départ le principe de *scaffolding* dans lequel la production langagière est conçue comme le produit d'un mimésis réussi. Il consiste donc, non pas à expliciter les règles de la grammaire ou de la rédaction, mais plutôt à attirer l'attention sur ce qui fait le texte ; pointer donc les composants propres à chaque type de texte, qu'il s'agisse d'un livre à lire ou d'une histoire à écrire. Pour ce faire, les spécialistes se sont beaucoup appuyés sur le principe de genre, que nous avons très peu explicité dans notre travail mais qui trouve tout son sens dans les principes de réalisation et d'instanciation (cf. tableau 8, figure 6 et figure 16 ci-après).

[56] Cf. Acevedo & Rose (2007) pour une présentation détaillée du programme que nous traduisons « Apprendre à lire : lire pour apprendre ». Le programme est basé en grande partie sur des concepts clés issus de la LSF.

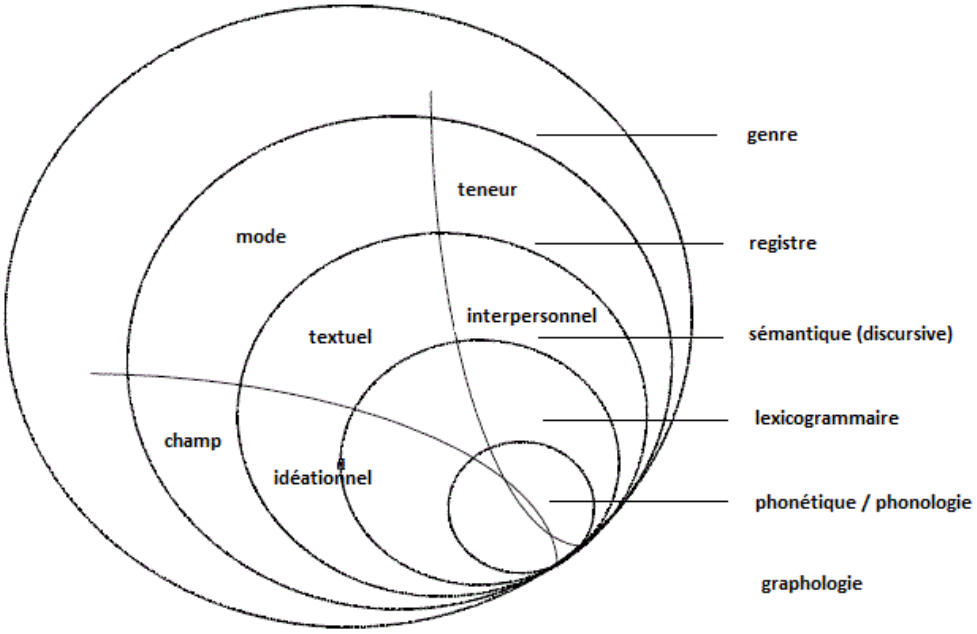


Figure 16 : La conceptualisation du genre, adapté de Martin (2003)

Par exemple, au lieu de chercher à remédier l'écriture en faisant appel à des ressources grammaticales (lexique et règles prescriptives de grammaire, entre autres), l'attention est portée sur le texte entier, comme répondant à un besoin précis. Donc, l'approche grammaticale traditionnelle est complétée par une approche ascendante, en aidant les 'apprenants' dès leur plus jeune âge à prendre conscience de certaines différences et exigences vis-à-vis de la rédaction comme répondant à un besoin précis : il nous incombe de préciser que beaucoup de cet enseignement se fait sans métalangage propre à la linguistique.

Outre ce programme particulier, nous relevons également de nombreux projets s'articulant autour de la problématique de la littéracie de manière générale, chez les apprenants de langue maternelle tous milieux et niveaux scolaires confondus. En effet, Christie (1991) retrace succinctement l'évolution des applications pratiques du cadre LSF aux enseignements de la littéracie en Australie, notamment à travers une exemplification des métafonctions expérientielle et textuelle. Dans cet exemple l'attention du scripteur est orientée vers le séquençage de l'information et donc sa distribution, le tout au profit d'une meilleure gestion de l'organisation du texte comme ensemble. De plus, l'auteur nous précise que cette approche qui peut paraitre simpliste a déjà porté ses fruits dans le contexte local depuis plus de 10 ans. Elle souligne le fait que l'approche se voit adaptée aux apprenants ; selon qu'ils sont à l'école primaire, secondaire ou notamment en milieu universitaire où l'inscription dans un genre textuel – ou dans une communauté discursive spécifique – est de mise.

En langue étrangère

De manière générale, nous pouvons soutenir que les principes d'application réalisés en langue maternelle sont transposables en langue étrangère, et vice versa. Ceci est mis en évidence, entre autres, par le nombre important d'études réalisées autours de la littéracie, tant en langue maternelle qu'en langues dites de spécialité en milieu universitaire (cf. Coffin & Donohue 2012 ; Drury 1991 ; Ventola & Mauranen 1991). Toutefois, il est intéressant de noter que le cadre LSF a été très souvent exploité en langue étrangère à des fins comparables à celles de la langue maternelle, mais par le biais d'approches tout à fait différentes.

Le cas de la métafonction interpersonnelle peut témoigner de ces approches ou « nouveaux terrains d'études ». En effet, cet aspect se voit souvent exploité en didactique, non seulement pour les ressources grammaticales qui permettent sa réalisation, mais plutôt en lien avec les rapports sociaux qu'il permet de manifester dans le discours ou le texte. Autrement dit, cet aspect facilite l'identification du positionnement du scripteur, à la fois face à son lecteur potentiel et à son texte. Ce point est d'autant plus pertinent dans ce contexte – en raison notamment du fait qu'un apprenant de langue étrangère peut très bien maîtriser la grammaire et avoir une connaissance lexicale tout à fait louable, sans pour autant que ses rédactions fassent émerger une véritable posture argumentative – ce qui pourrait être dommageable dans le milieu universitaire.

Hormis donc les ressources linguistiques que nous avons énumérées dans la section 3.2.5, concernant la métafonction interpersonnelle, cette particularité se fait à travers des études dites d'*Appraisal* dans la littérature anglo-saxonne. A ce sujet, nous pouvons citer Lancaster (2012), dans son travail de thèse portant sur le positionnement (qu'il nomme *Stance* et *Reader-Positioning*) chez des étudiants, qui affirme que la théorie dite d'*Appraisal* est un moyen d'approfondir la description de l'ensemble des systèmes de valeurs interpersonnelles réalisées dans un texte. Ceci n'est pas sans lien avec la notion de voix en L2 abordée dans le chapitre II, mais a le mérite d'avoir des composants clairement définis et identifiables selon des modalités établies dans le cadre systémique. A contrario, dans le chapitre II l'identification de la voix en L2 se fait tout d'abord sans critères stricts et passe donc par une phase que l'on pourrait qualifier d'interprétative. De façon succincte, Lancaster va jusqu'à nous préciser l'esquisse de son cadre d'analyse en soulignant le fait que cette valeur interpersonnelle s'analyse de la même manière que nos trois schémas d'analyse exposés en section 3.2.5, à savoir par le biais d'un système de réseaux.

> The framework makes use of three interrelated sub-systems to track speakers/writers' choices in interpersonal stancetaking: **Attitude** explores how feelings, judgments of

> people, and evaluations of things are built up in texts; **Graduation** explores how feelings and evaluations are subtly adjusted in terms of force and focus; and **Engagement** explores how "values are sourced and readers aligned" [...] through resources of modality, attribution, concession, intensification, and others. (Lancaster 2012 : 16)

Bien que celui-ci ne soit nullement confiné aux études didactiques, il permet dans le contexte précis d'apporter un autre regard sur le texte. Dans la mesure où la rédaction en langue étrangère, et dans notre cas précis en milieu universitaire doit répondre à des exigences en termes de genre (au sens de communauté discursive ayant une typologie textuelle et une structuration informationnelle strictes), l'*Appraisal* permet de prendre conscience de son positionnement en tant que scripteur notamment en rendant explicite les stratégies d'argumentation employées ; et ce, qu'elles soient bien ou mal exploitées. Notons cependant que les études en didactique tant en langue maternelle qu'en langue étrangère ne se limitent pas à des spécificités du cadre LSF. Nous avons voulu montrer par ces quelques exemples les apports de la linguistique systémique fonctionnelle dans l'évolution des tendances actuelles. Il convient à ce titre de signaler que tous les aspects de la langue qui sont traditionnellement étudiés par le biais d'autres théories peuvent l'être avec notre cadre. De plus, nous pouvons même affirmer, dans une certaine mesure, que certaines approches gagneraient à être mutualisées étant donné leur fort potentiel complémentaire.

3.3.2 En recherche et modélisation linguistique

La linguistique systémique fonctionnelle n'a pas seulement été digne d'intérêt et par conséquent n'a pas été exclusivement employée pour les questions ayant trait à l'enseignement des langues et l'analyse du discours. Elle a également été utilisée pour décrire les systèmes linguistiques de plusieurs langues en commençant par l'anglais, le chinois, le japonais et l'espagnol. La grande différence entre ces langues témoigne, à notre sens, à la fois de l'applicabilité et l'adaptabilité de la théorie à des systèmes linguistiques différents. De ce fait, lister des champs de recherche utilisés avec le modèle systémique nous semble une vaine entreprise – notamment parce que, comme nous l'avons soutenu tout au long de ce chapitre, la théorie se veut circonstanciée et robuste tout en restant modulable. Cependant, nous pouvons nous pencher sur son utilisation dans le domaine de la modélisation linguistique, qui n'est pas un domaine où beaucoup de théories du langage trouvent un champ d'application.

En effet, la LSF a également contribué de manière significative depuis les années 1970 à la modélisation et l'analyse des phénomènes linguistiques, notamment dans le domaine que l'on nomme aujourd'hui la linguistique computationnelle (cf. O'Donnell & Bateman 2005, pour une

présentation détaillée). Ceci est un aspect prégnant au vu du nombre croissant de logiciels, parseurs et autres applications utiles dans le traitement et la génération du langage naturel employant le cadre proprement systémique. Cependant, étant donné que ce point n'est pas développé davantage ni considéré pertinent dans notre analyse actuelle, il nous paraît ici judicieux de limiter la description de ces systèmes à une brève présentation de l'utilisation générale de LSF dans ce type de recherche.

A ce titre, nous pouvons renvoyer aux travaux de trois linguistes qui se sont penchés sur le côté computationnel de l'approche, à savoir (i) Matthiessen qui a contribué de manière considérable non seulement à l'approfondissement de la théorie générale mais également à la modélisation computationnelle (cf. Matthiessen 1983) ; (ii) Fawcett a opté pour une approche « formaliste » qui, admet-il lui-même, privilégie un aspect avant tout syntaxique et qui a mis au point des logiciels d'analyse linguistique utilisant un cadre systémique (cf. Fawcett 2004 ; Fawcett & Tucker 1990 ; Lin et al. 1993) ; (iii) Bateman et ses collègues ont également mis au point un système nommé KPML qui s'intéresse au traitement et à la génération du langage naturel (Bateman 1995, 1997 ; Bateman et al. 1999 ; Matthiessen & Bateman 1991) . Selon ces derniers, le cadre LSF s'avère particulièrement efficace pour le NLP/TALN. Bien entendu, ces trois auteurs sont cités à titre illustratif en raison notamment des différentes approches envisagées dans leurs recherches respectives, mais qui, d'un autre point de vue, ont permis l'élargissement à des concepts et domaines divers et variés. Notons enfin que le logiciel que nous utilisons pour réaliser l'ensemble de notre annotation et de l'analyse a également été développé par un systémiste, O'Donnell (2008, 2009). Celui-ci sera présenté plus en détail dans le chapitre suivant (cf. section 4.2.5).

3.4 L'apport de la LSF à notre étude

Aucune théorie n'est autosuffisante et chacune peut rencontrer de ce fait des limites. Il s'ensuit donc que toute étude entreprise dans une approche unique risque de se heurter à un moment ou un autre à ces mêmes limites théoriques : ceci est valable tant pour le cadre de la linguistique systémique fonctionnelle que pour les approches individuelles présentées dans les deux chapitres précédents. C'est en tout cas le constat que nous faisons à travers les différents cadres méthodologiques employés dans l'analyse des erreurs depuis 40 ans. En effet, les erreurs – qu'elles soient à propos des unités lexicales, phrastiques ou appartenant à des catégories supérieures – sont souvent envisagées à travers deux approches singulières ; d'une part, elles sont analysées à travers le prisme de la grammaire traditionnelle (cf. chapitre I) ; et d'autre part, dans une moindre mesure à

travers quelques avancées du cercle linguistique de Prague notamment pour ce qui est de la progression thématique et la structure informationnelle.

Toutefois, ce qui relève proprement du discours dans ces contextes précis ne bénéficie ni d'un cadrage ni de bases théoriques convenablement stables et solides. Cela étant, au vu du nombre croissant de langues étrangères enseignées en milieu universitaire et plus particulièrement dans notre cas de l'anglais de spécialité, il nous semble qu'une approche performante en analyse des erreurs pourrait permettre de relancer les études de ce type – qui à nos yeux n'ont ni suffisamment évolué ni su tirer profit des avancées en recherche linguistique. En effet, malgré le nombre significatif d'études réalisées sur la production en langue étrangère, deux constats s'imposent : les études sur l'analyse des erreurs s'inscrivent en grande partie dans une approche de grammaire traditionnelle d'une part et d'autre part uniquement certains types d'apprenants sont visés par ces études. En effet, certains semblent éviter d'étudier les textes d'apprenants qui n'auraient pas un niveau égal ou supérieur à C1 voire C2 ; c'est-à-dire une quasi-maîtrise en langue étrangère. Or dans notre expérience pédagogique, peu sont ceux qui y parviennent adéquatement. Cela est d'autant plus intriguant dans le contexte de la France où très peu d'études existent sur des corpus d'apprenants, indépendamment des différents niveaux de maîtrise. Et c'est justement tout l'intérêt de notre travail : à savoir répondre à la fois à un manque d'étude sur ce type de corpus et à un besoin d'incorporer un nouveau cadre linguistique, qui, a priori, serait susceptible d'apporter une perspective avant tout originale et prégnante.

De la syntaxe à la sémantique

Etant donné que l'ensemble des erreurs ne peuvent se résumer à des erreurs de grammaire en termes de règles de combinaisons morphosyntaxiques, il nous semble primordial de pouvoir identifier les différents niveaux d'analyse selon le niveau du système mise en cause, ou plus précisément selon ce qui n'aurait pas été maîtrisé. A priori une première distinction s'impose entre la syntaxe et la sémantique qui peuvent provoquer des erreurs de types différents. Cette distinction peut s'avérer significative dès lors que l'on souhaite avoir un regard holistique, tant sur les ressources lexico-grammaticales que sur le contenu sémantique qui a été visé. Dans cette perspective l'apprenant n'est plus un simple générateur de phrases, mais devient un créateur potentiel de sens. Il s'ensuite de ce fait que l'on doit examiner le sens qu'il a voulu créer et expliquer pourquoi il n'a pas réussi à l'actualiser de manière convenable.

Du local au textuel

Le passage d'une analyse « locale » à une analyse textuelle pourrait également permettre d'éclaircir davantage un phénomène qui n'est pas étranger aux enseignants de langue mais qui n'a pas encore fait l'objet d'études de manière substantielle et approfondie à grande échelle : à savoir la différenciation entre « erreurs locales » et « erreurs globales ou textuelles ». Ce phénomène constitue en effet une question épineuse que Carter-Thomas nous rapporte de la manière suivante :

> En tant qu'enseignant de langue je suis fréquemment amenée à juger de la qualité des travaux écrits de mes étudiants. Je parle ici d'une évaluation sur la réussite globale de ces écrits et de leur clarté. Or j'ai remarqué que, une fois corrigées dans un devoir écrit les erreurs syntaxiques, lexicales et orthographiques, l'ensemble reste souvent peu clair et décousu. (1999d)

D'après les principes d'instanciation et de la stratification (cf. figures 6 et 7) les trois types d'erreurs citées par Carter-Thomas relèveraient des instances lexico-grammaticales. Mais comme elle le souligne clairement, cela ne suffit pas pour garantir l'acceptabilité d'un ensemble textuel. Nous soutenons ainsi ici que l'on gagnerait à remonter l'axe de l'instanciation vers une catégorisation supérieure puisque les constituants de l'aspect lexico-grammatical ne sont pas les seuls éléments qui posent problème. Etant donné qu'on est dans la strate dit du *contenu* (*cf.* figure 5), une approche systémique voudrait que l'analyse se poursuive dans la strate sémantique. A ce titre, nous pouvons convenir que même si l'instanciation n'apporte pas de réponse en elle-même - la notion d'erreurs globales et locales pourrait être assimilée au rapport d'instanciation entre les constituants individuels d'un texte et son appréciation en tant qu'ensemble complet et cohérent. Autrement dit, l'erreur locale se situerait en fin de l'axe de l'instanciation vers l'*instance* tandis que l'erreur globale se rapprocherait plutôt du *potentiel* du système.

De plus, malgré l'apparente abstraction de ce concept, le principe d'instanciation peut s'avérer révélateur dans notre étude, dans la mesure où nous pouvons émettre des hypothèses sur les typologies et emplacements des erreurs, uniquement à partir du schéma de représentation issu de l'instanciation. En effet, en admettant que les erreurs identifiées dans notre corpus soient localisées sur une strate donnée, nous pourrons poursuivre en essayant de les expliquer vis-à-vis de l'axe d'instanciation. En adoptant une vision systémique, c'est-à-dire en examinant l'erreur tout d'abord « par le haut », l'acceptabilité de l'occurrence par rapport au contexte situationnel (en termes d'attentes textuelles ou de genre) pourrait être davantage explicitée. Ou encore en adoptant un regard « par le bas », il serait en principe possible de procéder à une étude sur les constituants en

termes de lexico-grammaire (c'est-à-dire l'exploitation et la manipulation correctes de l'ensemble des items fournis par le système global).

Enfin, si nous souhaitons dépasser les unités minimales, nous pouvons nous intéresser aux erreurs à des niveaux supérieurs à ceux traditionnellement étudiés de manière individuelle ou lexicale. Il serait désormais possible donc d'examiner l'ordre textuel et phraséologique, par exemple, par le biais d'un même cadre théorique. Il serait en principe également possible de s'éloigner des descriptions de surface qui constituaient une avancée majeure en 1967 ; c'est-à-dire les quatre premières distinctions de Corder – à savoir l'erreur dite (i) d'addition, (ii) d'omission, (iii) de sélection et (iv) d'ordre.

Cela étant dit, la possibilité d'interroger les erreurs qui seront annotées et présentées dans les chapitres V et VI selon qu'elles se trouvent dans le thème ou le rhème, dans le Mode, le Résidu ou à l'extérieur de ces deux derniers, nous intéresse tout particulièrement : sans oublier bien entendu les spécificités des erreurs portant sur la localisation, la position, la structure, la fonction non-maîtrisées, mais visées par les apprenants. Hormis donc l'étude approfondie des erreurs relevées dans notre corpus, notre objectif sous-jacent dans le présent travail est d'étudier dans quelle mesure le cadre LSF peut apporter des précisions sur les différents phénomènes cités ci-dessus et ceux qui seront mis au jour après l'annotation de notre corpus.

(Chapitre IV) Le cadre méthodologique

Dans les trois premiers chapitres, les principaux objectifs et postulats de cette étude ont été exposés, nous avons examiné des travaux antérieurs ayant des paramètres de recherche similaires aux nôtres et nous avons justifié l'orientation théorique adoptée dans l'étude actuelle. Nous passons maintenant aux considérations méthodologiques qui sous-tendent l'ensemble des résultats qui vont suivre. Dans un souci de clarté, ce chapitre a été divisé en trois sections distinctes afin de séparer ce qui relève proprement (i) de la réflexion méthodologique, (ii) de l'opérationnalisation des choix concrets et (iii) des tests de validité. En bref, notons que la réflexion méthodologique (ou la section 4.1) permet d'explorer les nombreuses considérations qui ont précédé la constitution du corpus ; la section 4.2 permet de présenter et justifier les choix concrets qui ont été réalisés ; et la section 4.3 permet d'interroger la notion de validité que l'on pourrait accorder aux différentes opérations d'annotations effectuées dans le présent travail.

Le besoin de cet abrégé méthodologique se justifie d'autant plus qu'il semble y avoir un écart significatif entre les différentes pratiques observées en termes d'exploitation des données portant sur des projets similaires selon (i) les années, (ii) le volume des données, (iii) et le cadre théorique. A première vue, cette disparité peut s'avérer normale ou anodine pour certains types de recherche mais pour des études en acquisition et plus précisément celles portant sur des corpus d'apprenants – une telle disparité pose la question de la reproductibilité et la représentativité, entre autres. A titre d'exemple, l'approche dite de/sur corpus n'a pas été celle privilégiée par de nombreuses études citées dans les chapitres I et II ou n'a été que partiellement exploitée par celles-ci. Or de

nombreuses études pionnières ont été réalisées à partir d'échantillons limités en nombre et donc insuffisamment représentatifs ou encore insuffisamment documentés : cela ne remet pas en cause les résultats obtenus mais cela nous permet de réfléchir à trois facteurs que les avancées méthodologiques en linguistique contemporaine ont permis d'optimiser : à savoir (i) la comparabilité de ces études avec d'autres ; (ii) la représentativité par rapport à la population étudiée ; et (iii) la généralisation à une population plus large. Ces points seront explicités davantage ci-dessous.

4.1 L'avènement informatique

Comme nous l'avons indiqué ci-dessus, cette première section du chapitre IV fournit des éclaircissements quant aux réflexions qui ont précédé le cadrage méthodologique de la section 4.2. Notons d'emblée que ces considérations sont largement admises par des linguistes travaillant sur des domaines d'étude précis (en synchronie ou en diachronie, par exemple), mais ne constituent pas encore à notre sens des pratiques très répandues dans notre contexte d'étude précis, notamment en France. Nous maintenons cependant que pour qu'une étude ou une expérience puisse revendiquer rigueur et solidité scientifique, il est important de respecter les procédures normalisées dans le domaine de l'étude. Cela étant, nous postulons ici que des choix éclairés issus des approches standardisées de la linguistique moderne – en termes de recueil et d'analyse de données - peuvent de manière générale constituer un cadre conceptuel substantiel. Cette normalisation a le mérite à la fois de concevoir des données qui pourraient être exploitées de manière discontinue par des chercheurs différents sans modifications supplémentaires et faciliteraient en principe la comparabilité des données sur le plan scientifique entre projets semblables.

Pour toutes ces raisons, les considérations méthodologiques ayant motivé nos choix et qui, de plus, sont issues de la linguistique de corpus sont succinctement décrites, de manière à positionner notre étude par rapport à d'autres études menées avec des paramètres linguistiques similaires. Il est tout d'abord question donc de présenter deux aspects méthodologiques complémentaires que la linguistique moderne a vu émerger en son sein ces dernières années ; à savoir la linguistique de corpus et la linguistique outillée. Nous retracerons ensuite succinctement le développement de ces deux approches, c'est-à-dire leurs principales contributions ; à la suite duquel nous décrirons une sélection d'applications, d'outils et de techniques qui ont été introduits dans la méthodologie de recherche linguistique en tant que conséquence directe de la recherche sur corpus. Enfin nous

expliquerons pourquoi et comment leurs applications peuvent être bénéfiques à notre cadre méthodologique.

4.1.1 Le cas de la linguistique de corpus

Dans notre contexte d'étude, rappelons que nous examinons principalement l'expression écrite en langue étrangère. De ce fait, qu'il s'agisse de focaliser sur la macrostructure textuelle, les éléments de la microstructure, les stratégies d'écriture mises en œuvre par les apprenants-scripteurs[57], ou encore d'identifier les stades d'acquisition atteints par le biais de leurs productions voire de repérer et analyser les erreurs produites - toutes ces questions ont un point de départ commun : à savoir le fait de travailler sur des textes réels, rédigés par de véritables apprenants de langue étrangère. Cela signifie que nous travaillons sur des textes qui relèvent de ce que l'on peut appeler du langage naturel (cf. Ellis & Barhuizen 2005). D'un point de vue didactique, mis à part le travail évaluatif, les enseignants travaillent également sur ces textes individuels et procèdent de manière générale à des comparaisons qui sont souvent limitées à des généralisations par rapport à une classe ou un groupe précis. Cependant d'un point de vue proprement de recherche, des chercheurs en linguistique appliquée effectuent ces mêmes comparaisons - entre autres - de manière plus approfondie soit de façon transversale soit longitudinale : ce qui signifie que plusieurs classes ou groupes sont comparés les uns aux les autres ou un groupe particulier est étudié à des intervalles de temps différents.

Il est important de noter ici que si les chercheurs souhaitent généraliser leurs résultats au-delà de leur échantillon ou même à une population plus grande que celle dont leur premier échantillon est issu - ils doivent choisir leur « échantillon ou population » d'une manière qui anticipe ce genre de comparaisons. Et c'est justement ici que la linguistique de corpus offre des procédures normalisées autour du recueil, du traitement et de l'analyse des données qui, lorsqu'elles sont suivies correctement, permettent d'assurer le caractère généralisable et la représentativité d'une étude, sans oublier bien entendu le fait de fournir des données à valeur doublement quantitative et qualitative. Mais on peut se demander légitimement ce qu'est la linguistique de corpus et comment elle a fait pour devenir une approche tant usitée, malgré les nombreuses réticences mises en avant pendant les années 1970 à 1990 ? On peut aussi se demander si elle est vraiment utile et qui l'utilise réellement. Pour répondre à ces questions, nous allons devoir retracer brièvement l'origine de ce que nous

[57] Eu égard à notre contexte d'écrits, nous nous référons de manière interchangeable à nos informateurs – en tant qu'apprenants, participants et notamment scripteurs.

considérons comme un ensemble de méthodes ou de procédures, et non une théorie du langage ou un champ d'étude à part entière comme peut le laisser entendre son appellation.

Pour tenter d'apporter des éléments de réponse, il convient de souligner que la linguistique de corpus - en tant que méthodologie en linguistique - est relativement jeune. Pour certains (cf. Leech 2011 ; McEnery & Wilson 2001), il existe deux périodes distinctes de l'histoire de la linguistique de corpus que l'on peut appeler pré-chomskienne et post-chomskienne. La première correspond aux années 1960 et témoigne des pratiques émergentes à la suite des travaux des structuralistes américains, de Randolph Quirk[58] et notamment de John Firth, l'un des précurseurs les plus attestés des études de corpus. Et la seconde renvoie à un regain d'intérêt au vu des théories alternatives naissantes et surtout des développements et de la démocratisation de l'accès aux nouvelles technologies vers la fin des années 1990 (cf. Léon 2005 ; Habert et al. 1997). Cependant sans entrer dans un vif débat sur l'émergence précise, l'identité du premier corpus construit ou la paternité anglo-saxonne du courant, intéressons-nous plutôt à ce qui fait consensus[59].

En effet, il est communément admis parmi les tenants de la linguistique de corpus (cf. Gries 2009 ; Hunston 2006 ; McEnery et al. 2006) que l'on peut attribuer au courant ces cinq principes définitoires simples. A savoir que la linguistique de corpus :

i. Renvoie à un ensemble de méthodologies sur le recueil, traitement (c'est-à-dire son équilibrage, étiquetage …) et analyse de corpus.
ii. Emploie des corpus issus de l'usage naturel de la langue, allant des œuvres littéraires, journalistiques, académiques aux discours du quotidien – sous des formes et médiums différents.
iii. S'intéresse aussi bien au quantitatif qu'au qualitatif mais avec une prédisposition notable pour les études de fréquence.
iv. Constitue une approche avant tout descriptive et neutre d'un point de vue théorique.
v. Permet le repérage et l'extraction de nombreux patrons (*patterns*) qui seraient passés inaperçus sans la perspective offerte par la linguistique outillée (concordancier, etc).

Face à la tradition linguistique dominante de l'époque, les approches sur corpus étaient marginalisées, au point où Léon (2005) nous rapporte qu'elles ont dormi pendant vingt ans et n'ont réapparu progressivement que dans les années 1980 - avec la montée en puissance des ordinateurs

[58] En dépit du fait que nous n'évoquerons pas davantage son nom, signalons simplement qu'il fut le fondateur du premier centre de recherche basée sur corpus, et fut également l'instigateur du célèbre Survey of English Use (SEU) initié en 1959 ou le London-Lund Corpus (LLC), son nom après l'informatisation.

[59] Notons qu'il existe des divergences vis-à-vis de l'émergence de la linguistique de corpus (date, raison, influence) A titre d'exemple Léon (2005) affirme que certains auteurs ont rétrospectivement construit leur propre histoire en exagérant ou en oubliant certains événements, faits et méthodes qui méritent d'être mis en perspective pour convenablement rétablir l'origine du courant.

facilitant la création et le partage de très grands corpus. Dans une certaine mesure, Williams (2006) et Habert et al. (1997) semblent rejoindre cette analyse – puisque pour ces derniers le regain d'intérêt de la linguistique de corpus est dû notamment aux nouvelles conditions d'accès à l'informatique. Toutefois, Habert et al. avancent deux autres explications possibles : (i) d'abord que l'engouement peut être attribué aux théories alternatives émergentes de l'époque, par rapport aux théories dominantes des années 1960 à 1990 et (ii) qu'il est potentiellement question de l'attrait des ressources créées par la linguistique de corpus.

> […] ces ressources sont désormais accessibles aux chercheurs universitaires pour des coûts raisonnables et ne sont plus réservées aux seuls centres de recherche industriels ou aux organismes qui ont constitué et mis au point ces données et ces outils. (1997 : 3)

A notre sens, les arguments avancés ne sont pas mutuellement exclusifs mais contribuent au contraire à apporter une vision d'ensemble. Aujourd'hui force est de constater que les applications de la linguistique de corpus ne se limitent pas à la linguistique au sens large, mais se voient appliquées à divers champs disciplinaires, à savoir par exemple, la sociologie et la psychologie sociale et à tout domaine où la langue en tant que discours constitue un objet d'étude. Cela signifie a priori que les approches issues de ce courant linguistique peuvent être appliquées à l'ensemble des domaines d'étude existant en sciences du langage, en passant bien entendu par l'ensemble des champs dits interdisciplinaires. Autrement dit, il est possible d'employer les approches de corpus dans les études allant de la pragmatique à la syntaxe ou de la linguistique appliquée à la linguistique cognitive. Nous reviendrons aux applications concrètes dans la section 4.1.3.

4.1.1.1 Recueil de données authentiques et comparables

Maintenant que nous avons dressé les contours généraux de la linguistique de corpus, il convient de nous intéresser à la question du corpus lui-même – puisqu'il constitue la matrice par excellence de la méthodologie. En effet, pour reprendre la définition de Sinclair (2005) « [a] corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. » Et comme le laisse entendre la définition, sa construction ne se résume pas à la simple collection des textes, mais répond à des exigences sous forme de ce qu'il appelle des « critères externes ». Pour Sinclair, il est de prime importance que l'ensemble du corpus soit choisi en lien direct avec la fonction communicative de la variété linguistique que l'on souhaite étudier. De plus, la variété doit être choisie au vu de son ancrage dans une situation de production réelle (critère externe). Ceci s'oppose au fait de choisir un corpus selon les spécificités linguistiques qu'un ou plusieurs texte(s) pourrai(en)t contenir (critère interne). D'autres critères externes sont par exemple le registre, le

mode (oral, écrit), le genre textuel, la variété linguistique (dialecte, aires géographiques ou culturelles …). En définitive, pour Sinclair les critères de sélection externes garantissent en quelque sorte l'authenticité du corpus.

Toutefois, notons un problème épineux auquel sont confrontés ceux qui construisent des corpus. En effet, l'un des soucis principaux - si ce n'est pas le plus important - pour ces derniers est celui de la représentativité. Comme le rapporte Leech (2006) la non-représentativité signifie que les résultats obtenus d'un corpus donné ne pourront en aucun cas être comparés ou généralisés à autre chose que le corpus duquel ils sont issus. Ceci constitue en soi un frein majeur, étant donné que la représentativité est de prime importance et représente souvent « un des arguments de vente » dans de nombreuses études sur corpus. La non-représentativité signifie alors que le corpus ne serait exploité que par ses concepteurs et est, de ce fait, voué à devenir ce que l'on appelle un « corpus fantôme » (Arbach & Ali 2013). A ce titre, nous rappelons également le constat alarmant d'Arbach & Ali qui révèlent que : « [f]orce est de constater que la plupart des corpus constitués en France n'ont servi qu'à leurs propriétaires » (ibid.).

Sans bien entendu dire que la non-représentativité est le seul facteur qui conduit à des corpus fantômes, nous soutenons que si la représentativité demeure bien assurée, documentée et facilement vérifiable, il se peut qu'il y ait davantage de suivi dans la promotion ou la vulgarisation continue de ces corpus : ce qui aurait pour effet d'empêcher ou de retarder l'arrivée de « l'état fantôme ». Nous reconnaissons toutefois que certains corpus de par leur degré de spécialisation peuvent n'être accessibles ou « intéressants » que pour un nombre restreint de scientifiques, ou encore le coût prohibitif peut également constituer un obstacle au partage entre chercheurs. Mais à notre sens, étant donné que la construction d'un corpus n'est pas en soi chose aisée et demande un investissement avéré, son abandon systématique après quelques études pose question. Cela nous amène aux trois questions suivantes : qu'est-ce vraiment la représentativité dans un corpus ? Qu'est-ce qu'un corpus comparable ? Et comment fait-on pour les assurer ?

La représentativité suppose que le corpus dont elle est le garant est constitué d'échantillons suffisamment représentatifs de la variété ou spécificité linguistique dont elle est porteuse. Par exemple, un corpus de français moderne supposerait que la langue orale et écrite seront représentées et les spécificités ou variétés connues de l'oral par rapport à l'époque, par exemple, feront de ce fait partie de ce sous-ensemble ; il en va de même pour des différents registres, genre et type d'oral – à savoir dialogue, entretien, émission de radio, et ainsi de suite qui doivent tous être pris en compte afin de garantir la représentativité du corpus. Signalons ici que la notion de

représentativité diffère selon Sinclair (2005) et Biber (1993) : le premier y voit un besoin de refléter la réalité d'usage effectif tandis que Biber prône ce qui s'apparente à une représentativité proportionnelle à l'ensemble des variations connues. Le but de cette « représentativité proportionnelle » est de mieux saisir la réalité variationnelle dans un échantillon d'une langue précise (cf. Arbach & Ali 2013 pour une discussion détaillée sur les différences d'interprétation chez ces deux linguistes). Précisons par ailleurs que nous rejoignons la position de Sinclair et, pour expliquer ce choix, intéressons-nous maintenant à notre cas de figure précis.

Notre corpus de travail, rappelons-le, est un corpus d'apprenants francophones rédigeant en anglais (cf. 4.2.1.3 pour une présentation détaillée). Avant de le compiler, il a fallu définir le type de corpus à établir[60] : à savoir entre un corpus dit embryonnaire, un corpus de référence, un corpus spécialisé et ainsi de suite. Dans notre cas, il est à la fois embryonnaire et spécialisé en ce qu'il n'a pas vocation à être largement diffusé pour incarner de manière générale l'ensemble de la population estudiantine apprenant une langue étrangère et d'un autre côté, il fait état d'une variété très précise de l'usage linguistique : à savoir des textes écrits, répondant à un seul genre textuel, avec un vocabulaire et jargon souvent propres au domaine de spécialité - le tout issu d'un groupe de locuteurs très restreint. La réflexion sur la représentativité se fait alors comme suit : (i) sur la population et (ii) sur le type de texte.

Pour ce qui est de la population d'apprenants étudiée : le corpus a vocation de représenter l'ensemble de la population estudiantine dans l'établissement de l'étude, à un niveau d'études précis. Donc les informateurs ou sujets-participants doivent incarner en quelque sorte la réalité effective de l'ensemble des étudiants visés. Ces réalités ou spécificités deviendront des variables à prendre en compte, le cas échéant : à savoir la réalité socio-économique, linguistique, le sexe, l'âge, et ainsi de suite. Par exemple, si la population globale des étudiants est constituée à 65% de femmes, avoir un pourcentage en-deçà ou au-dessus de ce chiffre de participantes risque de fausser les résultats en termes de représentativité (bien entendu, si le sexe constitue une variable importante de l'étude). Il en va de même s'il existe un pourcentage attesté d'étudiants étrangers, il est judicieux de ne pas avoir un pourcentage trop éloigné de la réalité effective.

Pour ce qui est des textes, le principe est le même. Ils doivent être en parfaite adéquation avec ce qui se fait ou ce qui est « attendu » d'ordinaire dans le contexte de leur réalisation. Ici il s'agit de textes argumentatifs portant sur des sujets à visée restreinte (socio-économique, dans notre cas),

[60] Cf. Sinclair (2005) et McEnery & Wilson (2001) pour une liste des différents types de corpus existants.

d'une longueur plus au moins contrôlée. La représentativité est jaugée ici en fonction du caractère « normal » du corpus. Il faut donc réfléchir et tenir compte de la présence d'éléments inaccoutumés. Les textes recueillis respectent-ils ce qui est « habituellement » accepté dans ce contexte ; y a-t-il des textes dont la longueur pose question : anormalement long ou anormalement court ? Qu'en est-il des textes qui ne respectent pas le genre demandé et par conséquent ne sont pas représentatifs de la variété linguistique étayée dans l'étude ?

Pour Sinclair (2005), ce type d'anomalies risque d'affecter les requêtes ultérieures du corpus, il faut donc être prudent avant de les intégrer ou les écarter. En toute connaissance de cause, il faut tenir compte de ce point dans l'analyse des résultats finaux. De même, soulignons que l'équilibre dans un corpus est important mais que l'homogénéisation en dépit de son caractère hautement pratique demeure superficielle. D'une certaine manière, l'équilibre et l'échantillonnage en tant que garants de la représentativité priment donc sur l'homogénéisation. Il convient de ce fait au chercheur de se prononcer sur l'acceptabilité d'un texte comme partie intégrante du genre ou type textuel dont jouit le corpus.

> As long as the choice of texts in a corpus still rests ultimately with the expertise and common sense of the linguist, it is appropriate for the linguist to use these skills to reject obviously odd or unusual texts. In any variety of a language there will be some texts - "rogue" texts - which stand out as radically different from the others in their putative category, and therefore unrepresentative of the variety on intuitive grounds. If they are included because of some high principle of objectivity, they are just wasted storage in the computer. (Sinclair : 2005)

En bref, le but de la représentativité est double. D'abord il s'agit de s'assurer que le corpus est authentique, dans la mesure où les données sont issues de conditions de production aussi bien réelle qu'attestée et qu'elles reflètent de manière fiable la réalité des participants. Ensuite il est question de s'assurer que les données sont comparables et interrogeables. Comparable avec des données de types semblables, recueillies et traitées plus au moins selon les mêmes procédures normalisées, ce qui signifie que les requêtes faites sur ordinateur pourraient aboutir à des éléments de comparaison. Nous rejoignons de ce fait Hasselgard & Johansson qui soutiennent que « corpora that are compiled according to the same design criteria […] lend themselves to comparative studies » (2011 : 37).

4.1.1.2 La « montée en puissance » des corpus d'apprenants

Le regain d'intérêt pour les études de corpus ne s'est pas limité, comme nous l'avons déjà souligné ci-dessus, à la linguistique descriptive et théorique. Il s'est progressivement introduit dans les études de linguistique appliquée : tant pour répondre au besoin naissant d'avoir des données réelles

que pour avoir des données bien plus conséquentes par rapport aux pratiques en vogue avant les années 1990. En effet, dans la mesure où la linguistique générale fait désormais appel à des corpus pour vérifier des hypothèses ou procéder à des observations de phénomènes linguistiques qui n'auraient pas été possibles sans corpus, la recherche en acquisition des langues étrangères s'est progressivement dirigée vers des textes rédigés par des apprenants afin d'observer de près cette variété linguistique que l'on appelle communément l'interlangue[61].

Toutefois, si les études sur corpus ont bénéficié d'un abandon généralisé du courant linguistique dominant de l'époque pour avancer et s'imposer comme méthodologie viable – sans bien entendu négliger l'essor des nombreux projets de collaborations interuniversitaires qui dépassaient souvent les frontières nationales dans les années 1960-1970 et 1980-1990[62] –, les corpus d'apprenants n'ont pas bénéficié d'un tel engouement[63]. Selon Hasselgard & Johansson (2011) leur envol a été inspiré dans une grande mesure par les travaux de Sylviane Granger. En effet, pour ces derniers, les travaux de Granger ont permis d'apporter et surtout de « vulgariser » une approche systématique – notamment en « militant » pour une approche qui dépasserait le faible volume de données sur lesquelles portaient les études des productions d'apprenants. A cela s'ajoute le fait que les données demeuraient en quelque sorte l'affaire d'un seul chercheur, ce qui poserait problème dans le contexte de recherche actuel[64].

Outre les travaux de Granger et la mise en place de ce qui est devenu aujourd'hui un des plus grand corpus d'apprenants, à savoir l'*International Corpus of Learrner English* (ICLE)[65], les corpus d'apprenants ont su profiter de toutes les avancées en linguistique de corpus : ce qui s'est traduit à la fois par une montée d'intérêt généralisée et par leur établissement en tant que champ d'étude riche et prometteur. A cela s'ajoute également le fait qu'il était désormais possible de construire des corpus d'apprenants plus volumineux, systématisés, équilibrés et représentatifs selon les standards normalisés. De plus, tout cela peut se faire de manière suffisamment rapide et bien moins laborieuse par rapport à la période précédant l'avènement de l'informatique. Soulignons enfin que cette montée d'intérêt s'est manifestée de manière globale dans certains enjeux linguistiques et didactiques que nous examinons ci-dessous.

---

[61] Comme nous l'avons souligné dans la section 1.3.2, ce terme renvoie au système linguistique en cours de développement chez l'apprenant d'une langue étrangère.
[62] Cf. McEnery & Wilson (2001) et Williams (2006) sur l'historique des premiers projets de corpus
[63] Cela ne signifie pas, bien entendu, une absence totale d'études sur corpus d'apprenants pendant cette période. (cf. le projet ARCTA ; Kübler 1995)
[64] Cf. les sections 4.1.2.3 et 4.3 au sujet de la validité des annotations et la notion d'accord inter-annotateur.
[65] Cf. les sections 4.2.1.1 et 4.2.1.2 pour plus d'information.

Tout d'abord, les enjeux linguistiques. Un des principaux apports des corpus d'apprenants réside dans le fait qu'ils apportent une nouvelle perspective dans un domaine en manque de renouveau, et où la théorisation semble avoir atteint un certain statu quo. Cette nouvelle perspective offerte par des données plus conséquentes (en volume et en quelque sorte plus faciles à analyser, au vu des nombreux outils d'analyse informatique disponibles) facilite la vérification des hypothèses avancées au sujet des différents processus et stades d'acquisition des langues étrangères. L'idée ici n'est pas de tout remettre en cause, mais de les contre-vérifier en procédant à des nouveaux tests et expériences notamment en raison du fait que les moyens pour le faire existent dorénavant. Cette idée rejoint celle avancée par Ellis & Barhuizen qui nous affirment que :

> By collecting and analysing samples of learner language, researchers can achieve the two goals of second language acquisition research (SLA) (Ellis 1994); (1) a description of the linguistics systems (i.e. the interlanguages) that the learners construct at different stages of development and (2) an explanation of the processes and factors involved in acquiring an L2. (2005 : 15)

Pour ces derniers, l'étude des productions d'apprenants est un moyen idéal pour mieux comprendre les avancées de la linguistique appliquée et plus singulièrement les études en acquisition. D'un autre côté, Adolphs & Lin (2011) voient dans l'exploitation de ces corpus une perspective descriptive mais également typologique dans la mesure où ces corpus sont envisagés en tant que variété linguistique nouvelle - avec une fonction communicative autre que celle de la langue maternelle.

> At the same time, learner corpora can be used as a basis for better descriptions of different varieties that emerge from communication between speakers who communicate in a language other than their first language. (loc.cit)

De toute évidence, les corpus d'apprenants peuvent apporter une contribution non négligeable à ce que nous connaissons actuellement sur l'acquisition d'une langue étrangère. Mais d'un point de vue purement méthodologique, l'influence de la linguistique de corpus peut également introduire une valeur quantitative importante dans ces études. En effet, au vu des volumes importants des corpus d'aujourd'hui, il est désormais possible de travailler davantage la représentativité et par conséquent augmenter le potentiel de généralisation.

Quant aux enjeux didactiques, il est à souligner qu'ils sont nombreux mais que nous nous limiterons ci-après aux trois principaux. Notons d'emblée que ces enjeux sont présentés séparément ici dans un souci de clarté, mais découlent en réalité les uns des autres. Par exemple, Hasselgard & Johansson soulignent que les corpus d'apprenants facilitent l'identification de trois micro-

phénomènes linguistiques : « [c]omparing data from learner corpus and a N[ative] S[peaker] corpus enables the researcher to identify overuse, underuse and misuse in the English of the learners » (2011 : 39). Cet usage renvoie à celui que nous jugeons être le principal d'un point de vue didactique. Il est tout d'abord question en effet d'identifier les anomalies ou les erreurs dans les productions des apprenants. Nos deux autres enjeux font suite à ce premier travail de repérage.

Une fois les anomalies relevées – dans le sens des « overuse, underuse and misuse » - le didacticien sur le terrain peut exploiter les informations obtenues d'un corpus d'apprenants pour mieux adapter son programme pédagogique. Il s'ensuit alors qu'une des conséquences directes de ces études est souvent la conception, adaptation ou modification des programmes pédagogiques, selon « l'analyse des besoins[66] » constatés chez les apprenants. Par exemple, si une fréquence significative d'erreurs est identifiée au sujet d'une construction grammaticale donnée – ce point de grammaire pourrait faire l'objet d'une mise au point ou davantage d'explications dans le prochain programme du cours. De même, si l'on remarque qu'un point longuement abordé dans un manuel de langue – au vu des résultats d'un corpus – ne pose aucun problème particulier pour les apprenants utilisant le manuel en question – il pourrait alors être question dans ce cas de figure précis de réduire ce point dans le prochain manuel de langue. Donc, les corpus ici peuvent apporter une vision à la fois évaluative et rectificative.

De surcroît, ces mêmes corpus permettent de créer des manuels ou des dictionnaires et d'autres outils pédagogiques, souvent à partir de zéro et donc sans « prédécesseurs ». Il convient simplement d'identifier les besoins des apprenants ou les problèmes à fréquence élevée afin que les concepteurs - au vu des interrogations faites sur les corpus – puissent construire leur matériel pédagogique en ciblant les lacunes ou points faibles des apprenants. A titre illustratif, McEnery et al. (2006) soulignent que le corpus d'apprenants appelé le *Longman's Learner Corpus* est connu pour avoir été à l'origine des nombreuses modifications apportées à leur collection de « *Learner dictionaries* » et méthodes de langue.

4.1.2 Le cas de la linguistique outillée : outils d'analyse et d'annotation

La linguistique de corpus a donné naissance à que ce l'on désigne de plus en plus aujourd'hui sous le nom de linguistique outillée. En quelques mots, elle renvoie au recours systématique qu'ont les

[66] Cette traduction de « *needs analysis* » renvoie aux enquêtes réalisées dans le but d'identifier les besoins d'un groupe d'apprenants en langue étrangère. Ces enquêtes visent principalement à recueillir des informations qualitatives des apprenants par le biais de questionnaires, entretiens, et ainsi de suite. Mais elles peuvent également avoir un pendant qualitatif, en s'intéressant tantôt aux résultats évaluatifs tantôt aux usages réels qu'ont fait les apprenants d'une langue étrangère.

linguistes aux ressources ou techniques informatiques ou informatisées, permettant d'exploiter de manière optimale l'ensemble des données d'un corpus. Notons également que l'utilisation de ces outils varie énormément, mais qu'elle renvoie à un phénomène qui est de plus en plus observé en linguistique descriptive. On pourrait même affirmer que tous les utilisateurs de corpus font appel à la linguistique outillée puisqu'il est tout simplement inconcevable d'analyser efficacement des corpus modernes – étant donné leur volume - manuellement. Quant à ceux qui ne travaillent pas sur corpus, une telle affirmation nous semble prématurée. Mais en dépit du fait que les outils existent et sont largement exploités par ceux qui utilisent des corpus, la linguistique outillée n'est souvent pas évoquée ou expliquée à sa juste valeur – aussi bien dans la littérature francophone qu'anglophone. Et à ce titre, nous reprenons deux citations de Laurence Anthony, le créateur d'un des concordanciers (cf. Anthony 2013a, 2013b) le plus en vogue en ce moment :

> However, one aspect of corpus linguistics that has been discussed far less to date is the importance of distinguishing between the corpus data and the corpus tools used to analyze that data. In any empirical field, be it physics, chemistry, biology, or corpus linguistics, it is essential that the researcher separates the actual data from the appearance of that data as seen through the observation tool. (2013a : 143)

> In corpus linguistics, on the other hand, researchers have tended to pay less attention to this separation. In fact, there is a continuing tendency within the field to ignore the tools of analysis and to consider the corpus data itself as an unchanging 'tool' that we use to directly observe new phenomenon in language. (loc.cit)

De même, nous jugeons judicieux de présenter brièvement les différentes options qui s'offrent aux linguistes travaillant sur corpus, du moins les options en termes d'outils d'analyses – et ce, d'un point de vue d'un non informaticien.

4.1.2.1 Les outils d'analyse

Il est important de signaler tout d'abord que les outils existants sont nombreux (cf. Poudat 2003 ; Pincemin 2009 ; Anthony 2013a) et que cette pléthore de choix peut avoir des inconvénients, ne serait-ce que le fait de devoir en choisir un parmi tant d'autres avec de nombreuses fonctionnalités quasi-identiques. Certains outils par contre varient selon leur disponibilité, leur niveau de spécificité, de spécialité ou complexité. Pour ce qui est de la disponibilité, certains sont uniquement disponibles en ligne et ne permettent que de faire des interrogations ou des requêtes ponctuelles, soit en utilisant un corpus fourni par le logiciel lui-même comme *CQPweb*, soit avec la possibilité d'incorporer un corpus « personnel » comme pour *SketchEngine*. Mais au-delà de ces outils qui demandent souvent une certaine initiation ne serait-ce qu'avec le jargon technique propre aux

logiciels, il y a d'autres logiciels en ligne qui, à première vue, sont moins performants mais accessibles à un public plus large (cf. par exemple *Anatext* v.2.3). Ensuite, ceux qui sont téléchargeables moyennant une contrepartie financière (allant de l'accessible au prix prohibitif) ou en libre téléchargement nous semblent – en tant que non informaticiens – bien plus accessibles, notamment dans la mesure où l'interface est souvent plus ergonomique et donc plus intuitif.

Vient ensuite leur spécificité : certains outils ont vraisemblablement été construits pour l'usage d'un seul corpus tandis que d'autres sont adaptables – notons qu'il est rare qu'un seul outil fournisse toutes les fonctionnalités dont on a besoin pour interroger de manière exhaustive un corpus donné. Enfin, d'autres outils se démarquent par leur simplicité d'usage et d'autres par leur complexité. En effet certains ont été créés par des linguistes informaticiens pour des linguistes informaticiens ou des initiés, tandis que d'autres sont accessibles avec une formation rapide. Ce dernier point constitue selon Anthony (2013a) une des raisons pour laquelle la linguistique outillée n'est pas traitée de manière systématique dans la littérature linguistique. En effet, beaucoup de ceux qui exploitent ces outils ne le font que d'une façon personnelle, selon le besoin d'un projet spécifique – et pour cause. S'ils ne sont pas eux-mêmes issus d'une formation informatique, ils peuvent trouver l'explication du fonctionnement quelque peu abstrait ou inaccessible aux non-initiés des algorithmes ou autres jargons informatiques.

Toutefois soulignons que malgré l'existence de nombreux outils dits de quatre générations différentes (cf. McEnery & Hardie 2012 ; Anthony 2013a), une grande majorité de linguistes ne semble pas consulter beaucoup de ces outils pour faire une comparaison exhaustive avant d'adopter un outil spécifique. Le choix semble se faire à minima – ou se résume souvent à ce que l'on connait ou ce qui est utilisé par ses collègues. Signalons à titre indicatif les principales différences de ces quatre générations d'outils dans l'ordre d'apparition.

(i) Les outils créés dans les années 1960 étaient utilisés pour analyser uniquement des corpus d'une taille restreinte. De plus, ils avaient des fonctionnalités minimes, se limitant souvent à créer des statistiques d'occurrences lexicales en termes de *key-word-in-context* (KWIC) et nécessitaient un ordinateur central ou un mainframe, ce qui nous donne toute de suite le sens de la mesure ; étant donné qu'un mainframe à l'époque était l'apanage de quelques-uns.

(ii) La deuxième génération a vu l'apparition de nouvelles fonctionnalités en termes de pré- et post-texte notamment dans un concordancier désormais amélioré. A cela s'ajoute le fait non-négligeable que l'on pouvait désormais travailler sur des ordinateurs de bureau,

ce qui a permis d'étendre les études sur corpus à une échelle individuelle. Cette période correspond à la naissance de la lexicométrie en France.

(iii) La troisième génération qui a commencé dans les années 1990 fait référence aux outils les plus utilisés de nos jours (cf. par exemple *AntConc, TXM*). Elle a facilité l'intégration des fonctionnalités déjà existantes avec de nombreuses méthodes statistiques nouvelles, sans oublier bien entendu la possibilité d'employer des corpus multilingues. Les outils de cette génération permettent également de travailler convenablement sur des corpus allant jusqu'à plusieurs dizaines de millions de mots (au-delà de 100 millions, les outils rencontrent des problèmes). On pourrait ajouter aussi que cette génération dans les courants proprement anglo-saxons correspond à la mue de la lexicométrie en France, en ce que l'on appelle désormais la *textométrie*. En effet, ce qui la différencie de l'ensemble des approches de la 3ème génération est le fait qu'« elle se caractérise notamment par certains calculs fondateurs, statistiques (les spécificités, les cooccurrences) ou non (les segments répétés, les concordances), et accorde une place fondamentale au « retour au texte » (bien outillé dans les logiciels) pour interpréter les unités (généralement des mots) sélectionnées par les calculs » (Pincemin 2011).

(iv) Les outils de la quatrième génération sont les moins répandus, mais résolument les plus performants. Cette génération[67] concerne des corpus de taille très significative, dépassant souvent les 100 millions de mots. Elle fournit donc les outils permettant de prendre en charge ce genre de volume, avec des fonctionnalités facilitant des interrogations rapides, tout en empêchant que l'on puisse consulter le corpus entier pour des raisons de copyright. Toutefois malgré ces avancées, Anthony (2013a) souligne ce qu'il considère être une sorte de « overkill » (comprendre, excès), dans la mesure où il existe désormais un certain nombre de paramètres obligatoires à prendre en compte avant de débuter la moindre analyse avec cette quatrième génération.

En définitive, ces outils – de la première à la quatrième génération - permettent de faire des interrogations ou requêtes sur des items lexicaux individuels ou des unités multi-mots. Ils permettent également de visualiser l'ensemble des résultats sous différentes formes, par exemple à partir de la deuxième génération il est devenu possible de voir l'item sur lequel porte la requête avec ce que l'on appelle aujourd'hui le « pré-texte » et le « post-texte » autour de l'occurrence recherchée. Certains outils fournissent automatiquement des statistiques d'ordre général, d'ordre

[67] *SketchEngine* et *Wmatrix* sont deux exemples de la quatrième génération, selon Anthony (2013a)

textuel, voire sur l'usage de la référence pronominale et les unités les plus fréquentes, sans interrogations supplémentaires. D'autres permettent de visualiser les dépendances syntaxiques selon certaines théories linguistiques. Les analyses factorielles, les chaines de référence ou bien d'autres structures grammaticales ne sont pas en reste. Toutefois, comme nous l'avons souligné, cette pléthore de choix peut s'avérer intimidante pour celui qui ne cherche qu'à faire des requêtes simples en termes de fréquence dans un petit corpus personnel et qui n'aurait pas eu d'initiation préalable aux jargons de bons nombres de ces logiciels.

4.1.2.2 Les outils d'annotation

Ce que nous appelons annotations, c'est l'ajout d'informations au corpus afin d'optimiser l'analyse ou selon la définition de Habert (2004) : « ce sont des programmes qui ajoutent par exemple à des corpus bruts ou déjà annotés de nouvelles informations comme le lemme, la catégorie morphosyntaxique des mots, les dépendances syntaxiques entre mots, des liens de coréférence, etc. ». De toute évidence, les annotations peuvent être de types très variés. En premier lieu, il peut être question d'ajouter des métadonnées sur les textes individuels qui constituent le corpus. Par exemple, il pourrait être question d'encoder des segments textuels selon les fonctions qu'ils occupent dans un texte – comme le cas d'un corpus d'entretiens. Il serait non-négligeable de pouvoir identifier la source des textes, d'avoir des informations sur les informateurs, d'avoir des informations sur les phénomènes extralinguistiques se produisant en même temps que l'entretien (rire, pause, etc.). Pour des textes écrits, il pourrait être question d'annoter les différentes sections à l'intérieur du texte, à savoir l'en-tête s'il y en a ou le titre par exemple – de façon à les distinguer des différentes parties du corps du texte (introduction, développement, conclusion). Si l'analyse porte sur la structure rhétorique, il pourrait être question de signaler les différents arguments ou prises de position dans le texte, et ainsi de suite.

Toutes ces informations sont souvent à ajouter manuellement ou de façon semi-automatique. Aujourd'hui de nombreux outils permettent l'incorporation d'un système d'annotation d'un autre logiciel, ce qui permet la continuité voire le partage des analyses entre utilisateurs de plateformes différentes. In fine, soulignons que l'annotation peut intervenir à plusieurs niveaux d'analyse linguistique allant de la prosodie, de la part-of-speech (POS) jusqu'à la pragmatique (Leech 1997). L'annotation peut, bien entendu, être appliquée à notre cas de figure – en annotant les différentes erreurs repérées dans notre corpus d'étude selon des niveaux d'analyse linguistiques différents. L'ensemble de ces niveaux d'analyses est mis en exergue dans la section 4.2.3.

4.1.2.3 La validité des annotations

Dès lors que l'on a réfléchi aux différents outils d'analyse et d'annotation que l'on souhaite appliquer à un corpus donné, il convient de réfléchir à la validité qu'auraient les items – une fois l'annotation effectuée. En effet, si cette réflexion se veut désormais partie intégrante du processus d'annotation dans de nombreux corpus, c'est notamment dans le but d'assurer l'exactitude, la reproductibilité des analyses et la généralisation des résultats obtenus à l'ensemble du corpus. Ce processus est nécessaire si l'on prend en compte le fait que l'annotation manuelle – comprise dans une certaine mesure en tant qu'étiquetage interprétatif et donc subjectif (Artstein & Poesio 2008 ; Bruce & Wiebe 1999) – peut comporter une marge d'erreurs qu'il convient d'anticiper et, dans la mesure du possible, de corriger.

A titre d'illustration, prenons l'exemple suivant. Si tout au long d'un exercice d'annotation on annote un même item lexical comme étant une erreur d'un type spécifique, la régularité de l'annotation peut laisser entendre que la valeur singulière de l'item annoté a été manifeste et que l'annotateur a su reconnaître cette constance. Toutefois, si ce même item lexical - ayant une valeur contextuelle invariable - se trouve annoté de plusieurs façons équivoques, c'est l'ensemble du schéma d'annotations, le guide de l'annotation ainsi que l'aptitude de l'annotateur face au travail d'annotation qui doivent être remis en cause. Pour éviter donc cet aspect fâcheux, il convient tout simplement de tester la validité des annotations. Cela étant, nous soulignons la définition fonctionnelle donnée à ce procédé par Karën Fort.

> Les accords inter- et intra-annotateur sont complémentaires de l'analyse de la conformité sur mini-référence (qui peut correspondre à un accord entre l'annotateur et les experts), en ce sens qu'ils permettent de valider l'assimilation de la formation et du guide de manière directe, et de vérifier la cohérence de l'annotation réalisée. (Fort 2012)

Autrement dit, il s'agit de s'assurer que les annotations sont cohérentes d'un point de vue strictement méthodologique. Cette cohérence se traduit de manière générale des deux façons suivantes : (i) l'accord inter-annotateur en termes de comparabilité entre un corpus identique annoté avec le même schéma d'annotation - pour les mêmes phénomènes caractériels - mais par deux ou plusieurs annotateurs et (ii) l'accord intra-annotateur, c'est-à-dire, la comparabilité de l'ensemble des annotations effectuées par un même annotateur.

Nous notons qu'un accord élevé entre annotateurs signifie que l'annotation dans son ensemble est fiable. De plus, s'il existe un guide ou manuel d'annotation un accord élevé laisse entendre que celui-ci a apporté des précisions claires quant aux valeurs des différents étiquetages possibles. Si le

taux d'accord est faible entre plusieurs annotateurs, cela peut s'expliquer en partie par des étiquetages non-clairement définis et donc sémantiquement ambigus. Signalons que Gut & Bayerl apportent une précision qui doit également être prise en compte lorsqu'il s'agit des taux d'accord entre annotateurs.

> It was shown that almost perfect agreement between annotators is possible, but that agreement is correlated to the complexity of the annotation task. The higher the number of different categories in an annotation scheme, the lower the agreement. (2004 : 567).

Quant à notre étude, nous affirmons que le calcul[68] de l'accord inter-annotateur peut s'avérer fort utile, en dehors de la notion de validité, dans l'interprétation des erreurs à tous les niveaux d'analyses que nous avons établis. Nous affirmons également qu'il peut permettre de comprendre dans quelle mesure l'idée que l'on se fait de « ce qui constitue une erreur » peut varier d'une personne à une autre. Pour ce faire, nous avons procédé à nos propres tests d'accord inter-annotateurs et les résultats sont présentés dans la section 4.3.

4.1.3 L'apport de ces deux branches complémentaires à notre analyse

Maintenant que les ressources de la linguistique de corpus et de la linguistique outillée ont été succinctement exposées, nous soulignons que la première nous renseigne sur la méthodologie du recueil des données afin de les rendre « analysables » selon des normes que nous avons parcourues ; tandis que la seconde nous fournit ce que nous appelons les véritables outils d'interrogations et d'exploitations nécessaires à l'analyse de l'ensemble de nos données.

4.2 Recueil et traitement du corpus

Les postulats de départ mentionnés notamment dans la section 0.1 de l'introduction orientent et limitent nos choix de données : celles-ci sont soumises principalement à des conditions d'authenticité et d'actualité. En effet, comme il a été souligné dans les chapitres précédents ces deux premières variables permettent une description linguistique résolument plus objective et précise à condition que les données n'aient pas subi de manipulation préalable et reflètent la réalité linguistique avant et au moment du recueil. Ainsi, pour les besoins de la présente étude nous avons opté pour la constitution d'un corpus d'apprenants issu du milieu universitaire dans laquelle nous sommes intervenus ; un corpus qui se veut embryonnaire et exploratoire et non un corpus de référence. Embryonnaire dans la mesure où son potentiel n'est qu'en partie exploitée

[68] Cf. Artstein & Poesio (2008) pour une présentation détaillée des différentes méthodes de calcul d'accord entre annotateur, avec des précisions sur la fiabilité et le parti pris de chaque méthode.

ici (cf. section 4.2.4), et exploratoire puisqu'il a été réalisé sans collaboration extérieure et sans vocation à la généralisation.

Dans cette deuxième partie du chapitre IV, nous continuons de passer donc en revue notre cadre méthodologique, en nous intéressant essentiellement aux multiples choix effectués tout au long de l'étude au niveau du recueil et du traitement des données. Nous commencerons par l'identification de quelques variables qui ont ostensiblement orienté la collecte des données et celles qui permettent la différenciation entre sujets-participants. Nous esquisserons ensuite les modalités du recueil du corpus, dans ses étapes préparatoires jusqu'au traitement informatique.

4.2.1 Le besoin d'un corpus propre : les étapes préparatoires

En dépit du fait que des corpus de type similaire existent et sont disponibles à la communauté universitaire, le choix de constituer notre propre corpus s'est imposé pour plusieurs raisons. Nous en détaillons les trois principales ci-après.

Une des variables indépendantes principales de notre étude est « le milieu universitaire français », et plus précisément celui de l'anglais de spécialité dans les sciences économiques. Or dans les corpus d'apprenants mis à disposition de la communauté scientifique, peu ont été réalisés dans des contextes francophones (cf. Kawecki 2009 ; le projet ARCTA 1994-1999). De plus, parmi ceux qui sont libres de droit et exploitables par tous, un grand nombre ne relève pas d'écrits universitaires argumentatifs en termes de genres textuels (cf. *Indianapolis Business Learner Corpus*) ou moins encore d'une « seule formation » mais plutôt d'un département plus large (cf. *Written Corpus of Learner English*), ou même de populations hétérogènes regroupées principalement selon le niveau d'étude ou la langue maternelle (cf. *Scientext).*

Limiter donc l'échantillon à une seule formation universitaire nous permet d'avoir une homogénéité représentative conduisant à une analyse plus exacte de cette sous-population. Une sous-population qui se distingue singulièrement par la mise en place de cours de langue différents des autres formations. En effet, le contexte d'étude est celui où les étudiants suivent des cours d'anglais de spécialité adaptés au besoin de leur formation principale[69]. Cette variable vise, entre autres, l'étude de la langue de spécialité, à la fois en tant qu'objet enseigné et objet maîtrisé chez nos sujets-participants. Elle permet par exemple d'apporter un regard sur la maîtrise de la terminologie, pour ne citer qu'une possibilité.

[69] Cf. section 4.2.1.3 pour une discussion détaillée des sujets-participants

Une seconde variable (*ici, dépendante*) non-négligeable est le contexte d'écriture qui a trait à l'authenticité recherchée. En effet, ceci s'explique par les rédactions contrôlées qui nous paraissent le plus à même de refléter la réalité de l'écrit universitaire, à savoir celles issues d'examen où l'écrit fait partie intégrante de la formation, et où celui-ci est assujetti à des règles de rédaction strictes (par exemple, sans l'aide de ressources documentaires ou électroniques). A cela s'exclut donc tout écrit fait sur demande d'un chercheur ou un travail fait à la maison où le sujet-participant aurait pu bénéficier d'une aide (cf. *Uppsala Student English Corpus*). Nous notons que certains corpus ne fournissent pas suffisamment d'éléments à ce propos, ou n'ont pas été faits dans un cadre contrôlé (cf. projet MeLLANGE). Ce qui peut d'emblée biaiser l'échantillon, dans la mesure où la variable dépendante contextuelle ne permet pas une comparabilité fiable entre individus d'une même population.

Passons maintenant à la question du partage des métadonnées. Si le contexte de rédaction est certes important, les informations recueillies sur les sujets-participants le sont d'autant plus. Celles-ci constituent une variable indépendante qui contribue également à la comparabilité des individus étudiés et par conséquent à la généralisation à une population plus large en fonction des caractéristiques partagées. Par exemple, les pratiques ou parcours langagiers, passés ou présents, ponctuels ou renforcés des participants peuvent se révéler être l'élément clé d'une comparaison fiable.

D'après ces premiers constats, la fiabilité de ces variables essentielles dans notre étude ne peut être assurée que si nous suivons chaque étape de la constitution du corpus de travail nous-mêmes. Et ce, afin de garantir et d'optimiser la précision de l'analyse vis-à-vis de la population étudiée.

4.2.1.1 Pré-enquête : préparation du terrain

Les vingt dernières années ont vu le nombre de corpus d'apprenants se multiplier de façon exponentielle. Il en existe en plusieurs versions : monolingue, bilingue et multilingue, avec des hypothèses et théories sous-jacentes motivant leur constitution aussi diverses et variées que les conclusions didactiques et linguistiques qui en sont tirées. A en croire des listes fournies entre autres par l'université d'Oxford *Text Achive* et les *learner corpora around the world*[70], créées par l'université de Louvain, sa réputation n'est pas à refaire. En effet, prenons par exemple cette dernière qui ne comporte pas moins de 74 corpus d'apprenants uniquement en langue anglaise, sans compter ceux réalisés dans d'autres langues. Nous notons cependant que seulement trois de ces

[70]Cf. URL: http://www.uclouvain.be/en-cecl-lcworld.html. Notons que cette liste de 2013 n'est pas exhaustive.

corpus sont réalisés avec des étudiants francophones apprenant l'anglais : *ARCTA, Scientext, A Learners' Corpus of Reading Texts* et *The ANGLISH corpus*. Les deux premiers étant les seuls qui traitent l'écrit.

Ce manque d'engouement ne se restreint pas dans la communauté universitaire française à l'anglais langue étrangère. Selon Kawecki (2009) « les corpus d'apprenants de français langue étrangère restent quant à eux peu nombreux et de taille relativement réduite ». En effet, seulement deux[71] sont répertoriés dans la liste de l'université de Louvain ; c'est-à-dire des corpus de français langue étrangère réalisés en France alors que nous dénombrons plusieurs hors métropole (cf. CEFLE, FRIDA, FLLOC, Dire Autrement …). Il est important cependant de souligner que cela peut s'expliquer par le fait que de nombreux chercheurs construisent des corpus d'apprenants embryonnaires de taille variable qui ne sont pas mis à disposition de la communauté scientifique plus large, en raison du fait d'un paramètre de non-comparabilité à une autre population que celle de laquelle le sous-corpus est issu.

Notre objectif est donc de créer un corpus d'apprenants d'anglais langue étrangère réalisé en France et répondant aux critères établis – qui, de plus, permettrait de procéder à des comparaisons avec nos résultats et ceux obtenus dans des corpus effectués hors métropole. De ce fait, nous avons trouvé judicieux d'employer la méthodologie développée par le corpus ICLE, qui comme nous l'avons déjà souligné dans la section 4.1.1.2, a été largement reprise pour la constitution de corpus d'apprenants par de nombreux chercheurs. Cette dernière stipule deux étapes obligatoires :

- Etablir un profil avec des renseignements sur l'historique langagier de chaque sujet-participant via un questionnaire. Ce profil est appelé « *learner profile* ».
- Recueillir les textes écrits issus de l'un des deux genres textuels suivants, à savoir des « *argumentative essay* » ou « *literature examination paper* ».

4.2.1.2 Questionnaire sur l'historique langagier : étude pilote et distribution

La méthodologie du questionnaire ICLE a donc constitué le point de départ du nôtre. En effet, cette méthodologie préconise un document court en trois étapes. En premier lieu, il s'agit de recueillir des métadonnées concernant chaque texte individuel et leur condition de rédaction (renseignés ici sont, par exemple, le type et titre du texte, le nombre de mots utilisés, si la rédaction a fait l'objet d'un chronométrage, si les scripteurs ont eu droit à des documents lors de la rédaction, etc.)

[71] Notons que ce chiffre ne représente que des corpus proposés à ICLE. Voir Hidden (2008) et Shaeffer-Lacroix (2009) pour deux exemples de corpus exploratoires sans vocation à une généralisation plus large.

Viennent ensuite les données linguistiques personnelles qui ont trait à la nationalité, le niveau de scolarisation, la ou les langues des parents, et celles utilisées tant au foyer qu'à l'école. Enfin, la troisième partie s'intéresse exclusivement aux autres langues étrangères apprises ou connues par le sujet-participant : ce dernier est invité à les énumérer, simplement par ordre de compétence.

Malgré l'étendue de ce questionnaire, il nous a paru judicieux de le modifier afin d'avoir des informations complémentaires sur des aspects qui ont un intérêt central pour notre étude : à savoir l'impact éventuel des différents contacts langagiers sur l'*output* en langue cible. En effet, ces informations constituent des données importantes permettant de croiser l'analyse linguistique des textes individuels de notre corpus avec le parcours langagier des individus étudiés. De ce fait, les études de Marian et al. (2007) et Li et al. (2006) ont apporté des précisions non-négligeables pour l'amélioration du questionnaire ICLE. Ces deux études ont mis en exergue une sorte d'inventaire des *Learner Profile* (LP). En effet, comme le soulignent Li et al. (2006) qui ont comparé 41 questionnaires, qu'ils soient appelés *language history questionnaire* (LHQ), Language Experience and Proficiency Questionnaire (LEAP-Q), ou Learner Profile, ces documents se chevauchent sur plusieurs points : une différence significative réside non pas dans les questions posées, mais le degré de précision de l'information demandée. Nous avons choisi donc d'apporter des précisions sur certaines « questions ouvertes » dans le modèle ICLE, et d'ajouter des questions qui portent sur une ou plusieurs variables importantes dans notre étude.

A titre d'exemple, là où le modèle ICLE préconise l'indication des langues étrangères apprises ou connues en ordre de compétence, nous avons jugé important de fournir un cadre plus strict. Et ce, afin de réduire le risque d'évaluation trop vague ou subjective dans l'absence d'attestation officielle. Nous avons donc fourni le cadre suivant :

6) Parlez-vous plus d'*une* langue ? Non Oui *[Entourez la bonne réponse]*

a) Si *oui*, pouvez-vous les énumérer en les classant, ainsi que les activités ci-dessous, selon l'échelle suivante :

Très rudimentaire	Rudimentaire	Moyen	Assez Bon	Bon	Très bon	Quasi-natif
1	2	3	4	5	6	7

Langues	Niveau global	Compréhension écrite (internet, livres)	Compréhension orale (radio, télé)	Expression écrite	Expression orale

Un deuxième exemple sera le cadre suivant où il est question de séjour dans un pays anglophone. Le modèle ICLE propose à nouveau des questions ouvertes : « *quand, où et pendant combien de temps ?* » Or les études sur l'impact de la motivation sur l'apprentissage d'une langue étrangère démontrent clairement une corrélation entre le motif du contact langagier et le résultat qui en découle (Dörnyei 1998). La question du pourquoi est apparue donc indispensable ici.

D'où notre reformulation :

14) Avez-vous déjà séjourné dans un pays anglophone ? Non Oui

Si *oui*, est-ce qu'en raison :

Motif	**Où**	**Durée**
D'un séjour de vacances ?		
D'un séjour linguistique ?		
D'une formation (en anglais) ?		
Autre, précisez ici :		

En somme, une différence entre le modèle ICLE et notre questionnaire final[72] réside dans le degré de précision visé, notamment où ICLE se base sur des questions courtes et ouvertes, nous avons privilégié un cadrage plus stricte avec des questions fermées afin d'harmoniser les réponses et de minimiser les réponses trop vagues. De plus, dans la mesure où toutes les possibilités envisageables ne pouvaient pas figurer dans le cadre, nous avons toutefois pensé à fournir un espace de rédaction libre de façon à permettre aux sujets-participants d'ajouter toute information supplémentaire qu'ils jugeraient importantes.

4.2.1.3 Les sujets-participants : groupe d'essai et sélection élargie

Pour constituer notre corpus, nous avons choisi de recueillir des écrits universitaires issus du département dans lequel nous sommes intervenu. Ce choix s'explique par le fait que ledit milieu avec son programme d'anglais de spécialité et le niveau général des étudiants ne nous étaient pas étrangers. Cela constituait plusieurs avantages notamment en terme de facilitateur d'accès à un nombre important d'étudiants, et donc indirectement à la taille du corpus. Nous nous sommes donc intéressé aux étudiants de première année, d'un même département[73]et de la même formation : à

[72]Cf. Annexe (A1) pour le questionnaire complet.
[73]Le programme de langues diffère selon les départements : celui duquel sont issus les participants est axé principalement sur le monde socio-économique.

savoir la Licence Sciences des organisations et des Marchés (LSO) de l'université Paris-Dauphine[74].

Le choix de l'année a été retenu par défaut puisque l'accès aux écrits des étudiants en année supérieure n'a pas été autorisé. Cela n'est pas pour autant un « obstacle » dans la mesure où la première année regroupe un plus grand nombre d'étudiants et représente une jauge en termes de comparabilité vu le nombre important de caractéristiques partagées. Par exemple, ils sont issus majoritairement d'une même année et série du baccalauréat, d'autant plus que l'admission en première année se base principalement sur celui-ci. D'emblée, nous estimons donc que le parcours et le niveau général de tous les étudiants sont intuitivement assez comparables. Ce qui n'est pas le cas avec les années supérieures, étant donné la diminution de caractéristiques partagées. En effet, des conditions différentes existent pour ceux qui auraient fait leur(s) première(s) année(s) ailleurs.

Les étudiants de première année ont donc été répartis en deux catégories. Ceux qui ont suivi nos propres cours d'anglais versus les autres. En effet, nous soulignons que les étudiants ayant suivi des cours avec nous ont été d'emblée exclus de la population finale, afin de ne pas biaiser les résultats. Cela s'explique par le fait que la problématique centrale de cette étude a ostensiblement orienté la manière dont nous avons enseigné ; et certaines questions centrales à la présente étude, à savoir la typologie des erreurs et leurs emplacements phrastiques, entre autres, ont été traités en cours, de façon à permettre aux étudiants de prendre conscience de ces phénomènes. Ces étudiants ont dont été retenus comme groupe d'étude pour tester les deux premières versions de nos questionnaires, avant distribution à une population plus large. En effet, les deux versions pilotes visaient, avant tout, l'identification d'éventuels problèmes non identifiés ou non-attendus. Elles nous ont permis de reformuler et clarifier des questions qui étaient restées sans réponses et d'autres où les résultats obtenus n'étaient pas exploitables en l'état.

Outre l'autorisation préalable de l'administration, il a fallu également l'aval individuel des enseignants pour contacter les sujets-participants. En effet, pour choisir les participants, plusieurs enseignants d'anglais des affaires qui interviennent en LSO ont été approchés afin d'expliquer le but et l'approche de la présente étude. Plusieurs ont refusé. Il nous a paru alors judicieux de rappeler davantage que les résultats ne porteraient pas de jugement sur leur travail individuel et seraient surtout anonymes, autant pour eux en tant qu'enseignants que pour les étudiants. Quatre ont donné leur aval pour contacter l'ensemble des étudiants sous leur responsabilité. Ces quatre

[74]Rappelons que l'admission se fait ici sur dossier et que 98% des admis ont reçu le baccalauréat avec mention : ce qui constitue une donnée à prendre en compte avant toute comparaison avec une autre population.

enseignants avaient un total de sept groupes d'anglais sur plus d'une vingtaine. Dans ces sept groupes, le nombre d'étudiants s'élevait à environ 180.

La distribution finale s'est faite directement par l'enseignant du groupe concerné, mais en notre présence afin d'optimiser le sérieux, le rendement et le nombre de participants. En effet, nous avons demandé aux quatre professeurs d'anglais de nous accorder 15 minutes pendant lesquelles les questionnaires ont été distribués, expliqués et recueillis. La version finale a été soumise à environ 180 étudiants. Nous soulignons que l'objectif précis de l'enquête n'a pas été précisé sur le questionnaire afin de ne pas biaiser les réponses. Il nous a semblé judicieux en effet de ne pas signaler aux sujets-participants potentiels que nous nous intéressions à leurs erreurs, et nous avons choisi par conséquent de mettre l'accent sur le cadre de notre travail de doctorat en sciences du langage et le laboratoire d'attachement.

4.2.1.4 Les copies d'examen : contexte de rédaction, collecte et tri

Il est à noter que l'identification positive de tous les signataires constitue une condition obligatoire autorisant l'accès aux véritables copies d'examen. De ce fait, dès lors que les questionnaires ont été rendus, il était tout d'abord question de vérifier que tous comportaient une signature lisible et avaient également été dûment remplis et datés, ce qui était une condition obligatoire pour l'identification positive des participants – sans quoi le questionnaire était considéré caduc. Cela dit, chaque signature nous a permis de consulter et recueillir les copies d'examen aussi bien du premier que du deuxième semestre du sujet-participant concerné. Par ailleurs, parmi les 180 questionnaires imprimés, 164 ont été rendus et 37 ont été tout de suite mis à l'écart pour signature ou identité inexploitable. Parmi les 127 signatures restantes qui nous ont permis d'accéder aux archives, cinq signatures de plus ont dû être écartées étant donné qu'une de leurs deux copies n'a pas été retrouvée. Ce qui nous ramène à un total de 122 signatures avec deux copies d'examen correspondant au premier et deuxième semestre, pour un total de 244 textes.

4.2.2 Traitement des données : numérisation, saisie de textes et anonymisation

Dès réception des 244 copies d'examen en version papier, il a fallu trouver une première solution de sauvegarde pérenne. Le choix de la numérisation a donc été retenu pour plusieurs raisons : tout d'abord en raison du fait que nous étions contraints de restituer les textes en version papier dans un délai limité ; deuxièmement, il fallait sauvegarder les éléments méta-textuels qui ne seraient pas retenus comme faisant partie intégrante du texte final, mais qui pourraient faire l'objet d'une étude ultérieure ou apporter des précisions nécessaires après l'analyse : à titre d'exemple, les brouillons

en marge, les ratures, les annotations et les évaluations des professeurs, les données permettant l'identification du sujet-participant, etc. L'intégralité des textes individuels a ainsi été numérisée, ce qui correspond à quatre feuilles doubles par sujet-participant, pour un total de 976 pages.

Notons que les copies d'examen comportaient deux sections : l'une portant sur un exercice de traduction d'un texte du français vers l'anglais et l'autre sur une dissertation argumentative avec trois sujets au choix. L'exercice de traduction se fait en cinq à huit lignes alors que la dissertation se fait généralement sur trois pages et demie, d'où le besoin de numériser la copie d'examen entière. Une fois la numérisation effectuée, il était question de dactylographier[75] les dissertations individuellement de sorte à les intégrer dans un fichier (*.txt*) : ce format étant le seul exploitable pour notre outil d'annotation[76].

Dans un premier temps, le nom et le groupe ont été saisis sur chaque copie afin de permettre une première sauvegarde authentifiable ultérieurement. Ensuite, ils ont été enlevés de sorte à anonymiser le corpus de façon complète. En effet, l'anonymisation vise, entre autres, à la garantie de l'anonymat et à la réduction de la subjectivité dans l'analyse, afin que l'identité du sujet-participant ou de l'enseignant ayant corrigé la copie ne puisse influencer le travail d'annotation.

4.2.3 Logiciels et schémas d'annotation d'erreurs

L'analyse des erreurs ne peut pas se faire de manière automatique. Selon Granger (2004) la tradition d'analyse des corpus d'apprenants demeure largement manuelle. Cela renvoie notamment au fait qu'en dépit de la recrudescence des logiciels permettant des analyses textuelles automatiques, peu sont compatibles avec des corpus d'apprenants. Cette dernière soutient que ces logiciels :

> […] - whether lemmatizers, part-of-speech (POS) taggers, or parsers – have been trained on the basis of native speaker corpora, and there is no guarantee that they will perform as accurately when confronted with learner data. (2004 : 128).

Si l'analyse est purement syntaxique, certains logiciels proposent une analyse semi-automatique qui nécessite un balayage complet pour relever les inexactitudes ou les phénomènes écartés. Cependant dès lors que l'on s'intéresse aux phénomènes sémantiques ou pragmatiques, l'analyse doit se faire

[75] Notons que dactylographier l'ensemble a été retenu comme solution après avoir obtenu plusieurs rendements non-exploitables en l'état, avec des outils de reconnaissance optique de caractères (OCR).
[76] Cf. section 4.2.3.1 pour une discussion détaillée sur le logiciel d'annotation employé

manuellement. En effet, ce type d'analyse nécessite un repérage individuel des occurrences à analyser avant de procéder à une annotation manuelle, ce qui est bien notre cas de figure.

De plus, l'utilisation des schémas d'annotation, créés à partir d'une taxonomie d'erreurs préalablement établie, devient indispensable pour celui qui souhaite analyser un corpus d'apprenants de taille conséquente. Et pour cause, ces schémas facilitent l'analyse à mesure que ces derniers sont intégrés à un logiciel permettant de visualiser simultanément aussi bien l'ensemble textuel à analyser que les différents choix d'étiquetages disponibles. De plus, ces analyses sont de plus en plus informatisées et génèrent une fois l'annotation terminée des données statistiques sur les occurrences annotées.

Ci-dessous nous allons présenter l'outil qui a été utilisé pour annoter notre corpus, avant d'examiner les différents schémas d'annotation qui ont été utilisés pour l'analyse finale.

4.2.3.1 Le logiciel d'annotation d'UAM CorpusTool

Pour l'annotation de notre corpus, nous avons eu recours à un logiciel[77] qui propose une interface propice à une annotation multicouches et multi-niveaux : « multicouche » dans le sens où il permet d'avoir plusieurs schémas d'annotation dans un même projet d'annotation et « multi-niveau » dans le sens où il n'y a pas de limite au nombre de niveaux de profondeur[78] qu'un schéma précis peut avoir. En effet, étant donné que notre approche vise notamment la comparaison de plusieurs schémas d'annotation existants (tout particulièrement le modèle d'ICLE et le modèle fonctionnel dit d'UAM CorpusTool) avant de procéder à l'élaboration de notre modèle expérimental, il nous a paru judicieux de nous armer d'un logiciel qui se prête commodément à ce type d'exploitation : à savoir (i) facilitant la création d'une arborescence très développée et (ii) permettant l'intégration de plusieurs niveaux ou schémas d'annotation dans un même espace de travail. Ces besoins spécifiques expliquent donc le choix de l'UAM CorpusTool.

[77] Notons, à titre d'information, que le choix final de logiciel se posait entre Analec et UAM CorpusTool. Et bien que les deux correspondent, d'une manière générale, aux outils de la troisième génération d'annotation et que le premier (Analec) ait été créé dans notre laboratoire de recherche (cf. Landragin et al. 2012), nous avons privilégié le deuxième (UAM CorpusTool) pour des raisons purement pratiques. Ces raisons sont détaillées ci-dessous.

[78] Cf. les différents schémas d'annotations dans cette section avec des niveaux de profondeur différents. Par exemple, les figures 18 et 19 ont un seul niveau de profondeur tandis que 20 et 21 en ont deux.

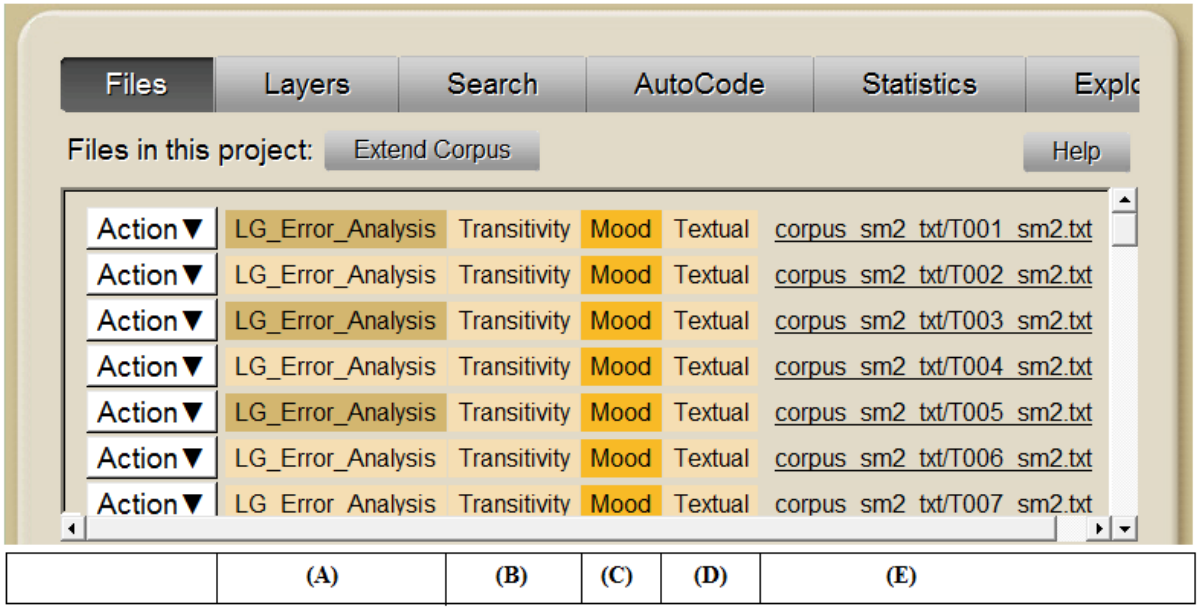


Figure 17 : L'interface d'UAM CorpusTool illustrant des schémas d'annotation exploités dans notre étude

Les quatre premières lettres (A-D) correspondent aux schémas d'annotation utilisés dans l'étude et la lettre (E) correspond aux textes individuels de notre corpus. Les différents schémas utilisés sont présentés ci-dessous.

Le logiciel UAM CorpusTool est l'évolution de deux programmes successifs et se définit selon son auteur comme « a text annotation tool primarily aimed at the linguist or computational linguist who does not program, and would rather spend their time annotating text than learning how to use the system » (O'Donnell, 2008 : 13). De plus, l'outil permet l'incorporation de plusieurs textes différents et permet également à l'annotateur de travailler sur l'ensemble des textes et des schémas d'annotations en même temps. En ce qui concerne les autres choix de fonctionnalités proposées, l'annotateur a le choix entre l'annotation d'un document entier, des segments textuels, des phrases, des syntagmes ou des items lexicaux individuels. Cela dit, le logiciel fournit des calculs statistiques automatiques tant comparatifs que descriptifs de l'ensemble des éléments annotés. Tous ces points ont motivé notre choix de logiciel, sans oublier son aspect intuitif ou « *user-friendly* » qui n'est pas négligeable pour nous en tant que non informaticiens.

Outre les raisons techniques qui ont guidé le choix du logiciel, le cadre théorique utilisé par son auteur a également été pris en compte. En effet, l'auteur a utilisé le cadre de la linguistique systémique fonctionnelle pour réaliser à la fois son arborescence (en termes de système de choix) et taxonomie des erreurs en langue anglaise. Le schéma résultant est le fruit de plusieurs études de corpus d'apprenants et constitue à notre sens une des taxonomies la plus élaborée à ce jour. En effet, notons que le modèle d'UAM semble avoir été conçu comme une alternative aux schémas

d'annotation proposés pendant plusieurs années et adaptés par de nombreux linguistes de corpus utilisant le modèle ICLE.

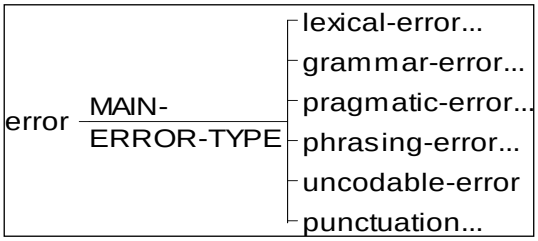


Figure 18 : La structure de base du modèle d'UAM

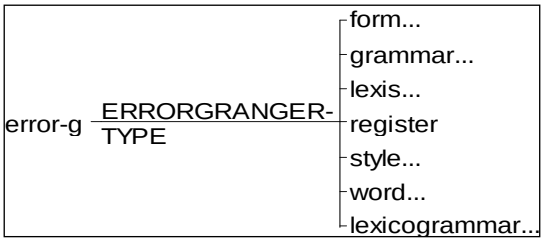


Figure 19 : La structure de base du modèle d'ICLE

A titre d'exemple, les figures 18 et 19 mettent en avant les deux modèles que nous comparons. Le modèle ICLE est indéniablement l'un des modèles le plus exploité, à un croire le nombre croissant de corpus d'apprenants réalisés et analysés avec cette même méthodologie. Toutefois, un regard avisé fera d'emblée ressortir l'approche de la grammaire traditionnelle dans l'élaboration de cette taxonomie. En effet, dans la structure de base[79] l'analyse semble partir du mot en tant qu'item isolé. Et aucune unité au-delà du mot n'est prise en charge par cette approche. A titre d'information « *Form* » renvoie simplement aux erreurs de morphologie tandis que « *Grammar* » renvoie aux classes grammaticales. On pourrait se demander par exemple quelle est la distinction faite entre « lexico-grammar » et « grammar » ou « lexis » et « word », qui à première vue nous parait bien arbitraire. Si l'on regarde la figure 20, ces questions restent en suspens.

Le modèle d'UAM prévoit dès le début, quant à lui, un cadre susceptible de signaler les erreurs distinctement en dehors de leur fonction grammaticale initiale. Ainsi l'annotateur a le choix entre l'annotation d'une erreur en fonction d'un point syntaxique, lexical, pragmatique voire au niveau proprement phraséologique. Cela étant, les catégories sont clairement délimitées voire parfois un

[79] Hormis l'approche basée sur une orientation proprement systémique (LSF), signalons que d'autres facteurs ont motivé notre choix du schéma d'annotation d'UAM CorpusTool. Par exemple, au début de notre recherche nous n'avions pas accès à la taxonomie d'erreurs complètes d'ICLE, avec les 53 étiquetages différents. De plus, l'accès au guide d'annotation et au logiciel propre à ICLE n'étaient pas en libre circulation. A titre illustratif, le schéma d'UAM compte plus de 130 étiquetages possibles, ce qui implique d'emblée une annotation plus fine.

peu trop. Cela devient évident si l'on compare le deuxième niveau de profondeur des deux modèles.

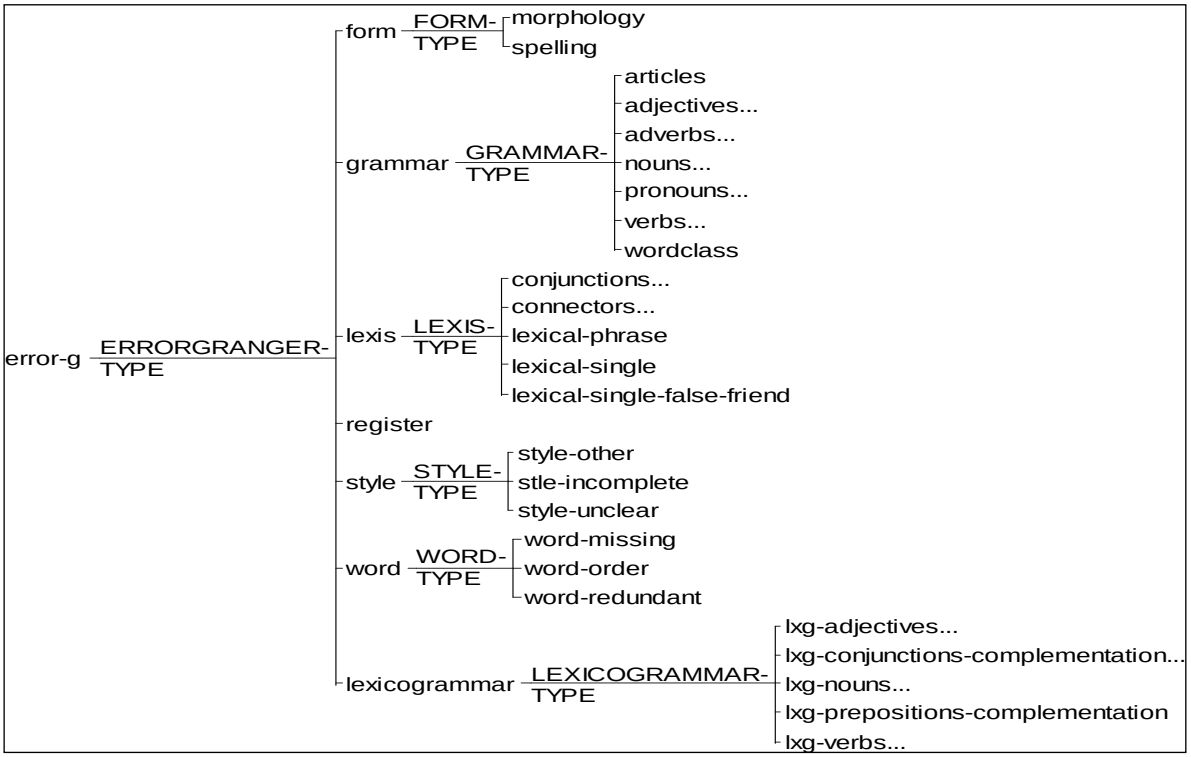


Figure 20 : Le 2ème niveau de profondeur du modèle ICLE (intégré à l'UAM CorpusTool, v.2.8)

Aux questions que nous venons de poser ci-dessus s'ajoutent des questions sur le sens de l'étiquetage « style » et l'ensemble de ses sous-groupes ou encore la distinction entre les six premières sous-catégories d'erreurs dans « Grammar » qui relèvent de ce que l'on appelle communément des classes ou catégories grammaticales et la septième sous-catégorie qui est intitulée « word class » (comprendre, classes grammaticales). Reconnaissons tout de même qu'un accès au manuel d'annotation aurait peut-être permis d'apporter de premiers éléments de réponse à ces questions.

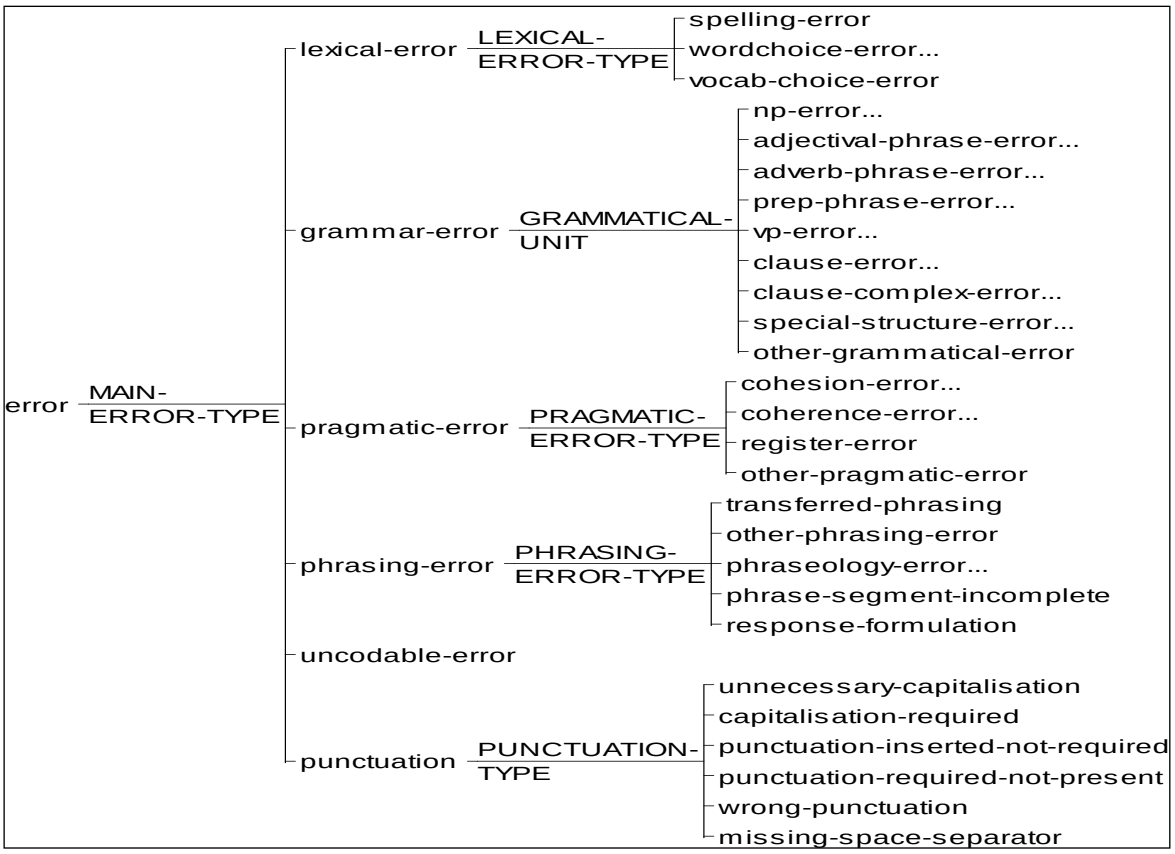


Figure 21 : Le 2ème niveau de profondeur du modèle d'UAM

L'idée de cette comparaison n'est pas de soutenir ici qu'un des modèles est plus performant ou mieux adapté à l'analyse des erreurs que l'autre, car nous pensons que les deux aboutiront potentiellement aux mêmes résultats. Mais force est de constater que le modèle d'UAM se prête plus facilement à l'analyse, dans la mesure où il fournit des catégories qui, à notre sens, sont clairement compréhensibles sans que l'annotateur ait besoin du manuel d'annotation ou des précisions sur les différentes délimitations des erreurs. Notons également à titre accessoire que le modèle d'UAM comporte jusqu'à quatre niveaux de profondeur.

4.2.3.2 Les modèles exploratoires issus des métafonctions LSF

Les trois schémas d'annotation ci-dessous sont à mettre en rapport direct avec les métafonctions qui sont amplement décrites dans le chapitre III, section 3.2.5. Cela étant, nous les présentons ici à titre purement illustratif. Précisons que les schémas LSF serviront lors de la deuxième étape d'annotation des erreurs : c'est-à-dire toutes les occurrences erronées relevées dans le corpus seront tout d'abord annotées avec le schéma d'UAM pour ensuite être ré-annotés par les schémas LSF[80].

[80] Comme nous verrons dans les chapitres V et VI, les erreurs sont annotées avec plusieurs schémas d'annotation différents. À titre d'information, dans le chapitre V, les erreurs sont tout d'abord annotées avec le schéma d'erreur d'UAM CorpusTool et ne sont ré-annotées ensuite qu'avec deux schémas LSF (parmi les trois disponibles) : à savoir le

Le but de cette ré-annotation est de comprendre dans quelle mesure ces métafonctions peuvent nous aider à mieux cerner les erreurs des apprenants, en nous focalisant à travers ces ré-annotations sur le sens que les apprenants cherchent à créer et non strictement sur les items grammaticaux non-maîtrisés.

Le niveau interpersonnel

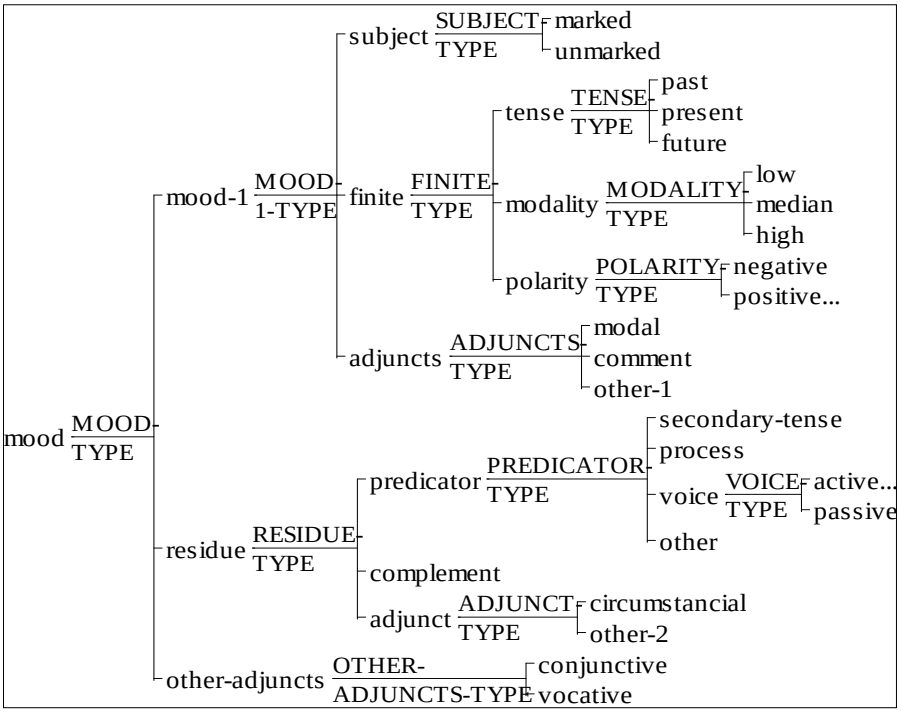


Figure 22 : Le schéma d'annotation n°2 (niveau interpersonnel)

Le niveau textuel

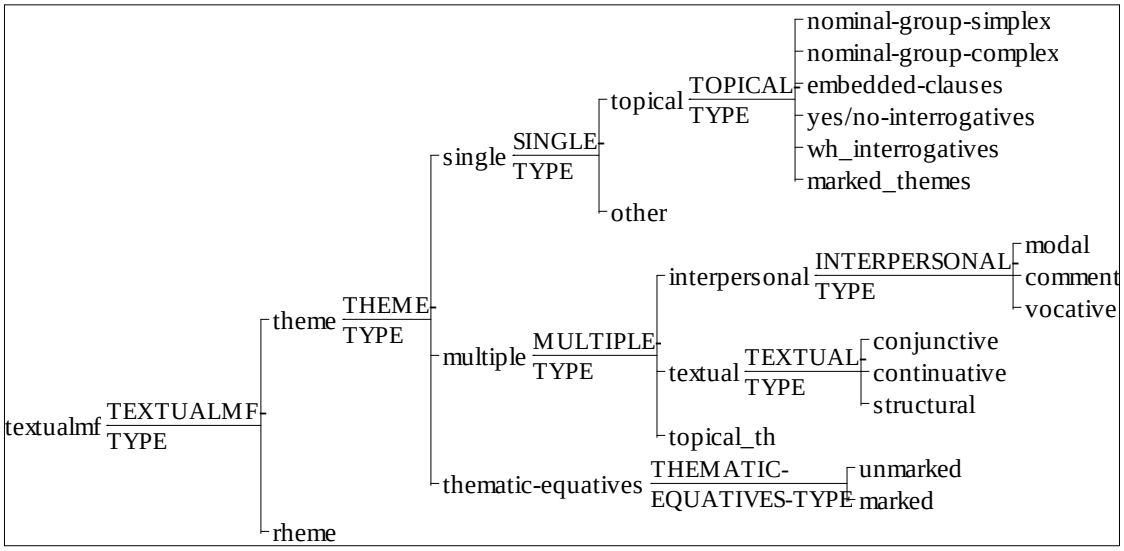


Figure 23 : Le schéma d'annotation n°3 (niveau textuel)

---

schéma expérientiel et le schéma textuel. Nous avons en effet jugé judicieux de ne pas faire une ré-annotation, à ce niveau, avec le schéma interpersonnel, étant donné que les sous-catégories du schéma interpersonnel avaient été "plus au moins" intégrées dans les erreurs grammaticales du schéma initial d'UAM. (Re)faire donc une (ré-)annotation, séparément, avec ce schéma nous semblait donc redondant. Par contre la deuxième étape de la ré-annotation dans le chapitre VI (avec les erreurs textuelles) seront ré-annotées avec les trois schémas LSF.

Le niveau expérientiel

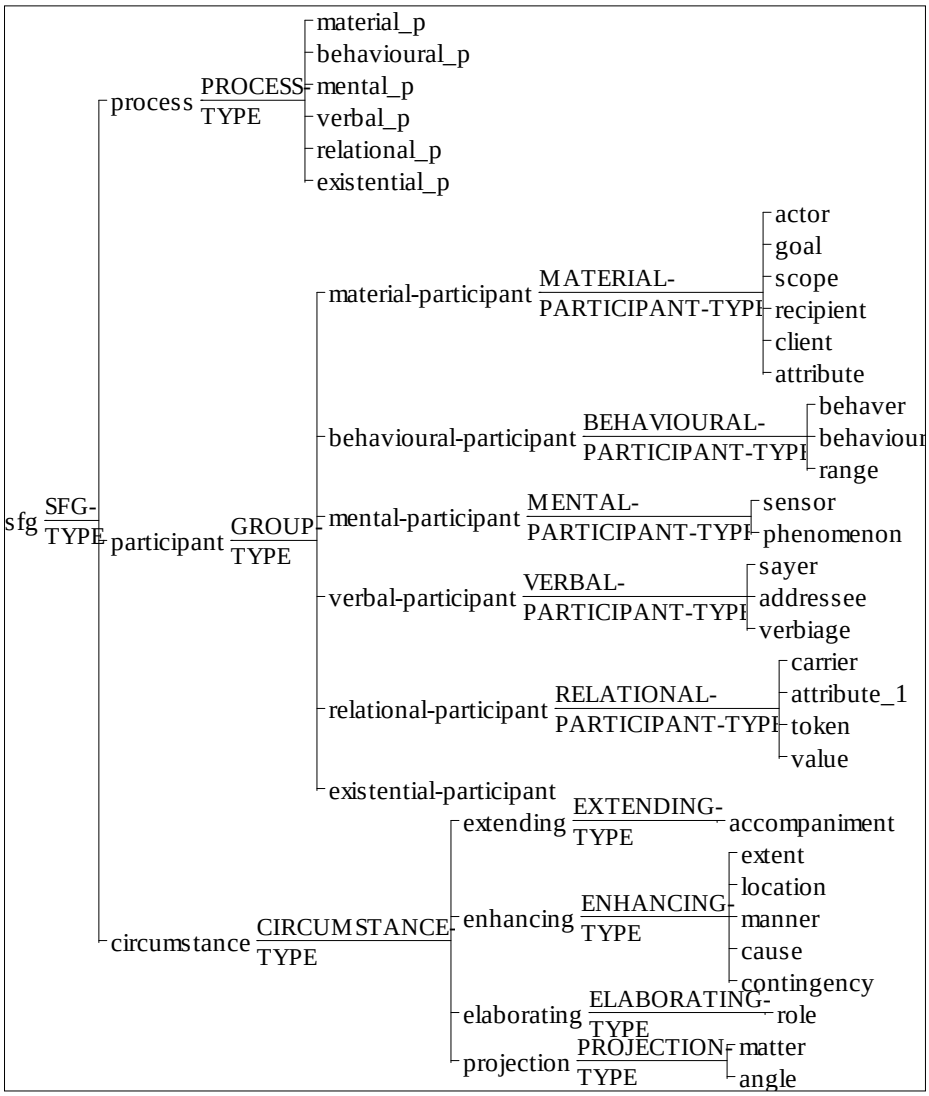


Figure 24 : Le schéma d'annotation n°1 (niveau expérientiel)

4.2.4 La répartition du corpus final en quatre sous-ensembles

Après avoir passé en revue les différentes considérations méthodologiques et théoriques qui ont précédé la constitution du corpus, nous allons succinctement expliquer comment ce dernier sera exploité dans la présente étude. En effet, comme nous l'avons signalé, le nombre de sujets-participants retenus pour l'étude s'élève à 122 et chacun nous a fourni deux textes : un au premier semestre (sm1) et un autre à la fin du deuxième semestre (sm2). Ce qui fait un total de 244 textes, pour 117 000 mots.

Erreur du système linguistique (volet 1)	Erreur d'acceptabilité textuelle (volet 2)	
txt 1 à 61 (sm1)	txt 62 à 122 (sm1)	122
txt 1 à 61 (sm2)	txt 62 à 122 (sm2)	122
122	122	244

Tableau 11 : les sous-ensembles du corpus

De manière rétroactive et dans un souci d'objectivité, il a été décidé de ne pas annoter les erreurs que nous appelons par la suite les erreurs du système linguistique (c'est-à-dire, en suivant le schéma d'UAM) et les erreurs de l'acceptabilité textuelle (cf. chapitre VI) chez un même participant de façon à éviter une sorte de double peine tant pour l'apprenant que pour l'annotateur. En effet, les productions des participants sp_01 à sp61, qui sont les mêmes au premier et deuxième semestre, seront analysées pour les erreurs du premier type tandis que les écrits des participants sp62 à 122 seront analysés pour des erreurs proprement textuelles. Signalons à titre d'information que l'ensemble des erreurs relevées dans les deux sous-groupes seront ré-annotées avec les schémas LSF.

4.3 Les test d'accord inter-annotateurs

Cette troisième sous-section fait écho à la section 4.1.2.3 dans laquelle la question de la validité des annotations a été abordée. Cela étant, nous détaillons dans les lignes qui suivent la procédure et les résultats des différents tests de validité – à savoir ceux permettant de vérifier la solidité et par conséquent la reproductibilité de nos annotations.

4.3.1 Le degré de fiabilité des annotations : tests d'accord inter-annotateurs

Une fois que l'étape de l'enquête-questionnaire[81] a été achevée et que le corpus a été recueilli, nous avons procédé à une annotation manuelle de l'ensemble de nos données. Cela dit, nous sommes conscient que toute annotation manuelle comporte une part de subjectivité et, de ce fait, une marge d'erreur humaine. C'est donc pour ces raisons que plusieurs tests de validité ont été mis en place pour déterminer le degré de fiabilité de nos annotations et conséquemment nous indiquer dans quelle mesure les résultats obtenus sont généralisables. De surcroit, étant donné que nous avons affaire à de multiples schémas et volets d'annotations, nous avons procédé à des tests de validité pour chaque schéma et surtout chaque volet d'annotation : et ce, à hauteur de 10% du corpus total.

Les tests ont été effectués par étape, avec cinq annotateurs (codeurs) différents. Ces derniers sont signalés R1, R2, R3, R4 et R5 et ne sont pas tous intervenus sur l'ensemble des tests de validité. En effet, ils ne sont intervenus que dans des tests où leur domaine de compétence constituait des paramètres que l'on voulait tester et comparer entre codeurs. A titre d'information, les profils succincts des annotateurs sont précisés ci-après :

[81] Cette étape est précisée davantage dans le chapitre suivant.

- R1 : nous-même
- R2 : enseignant anglophone, sans formation linguistique, avec plus de 15 ans d'expérience en anglais de spécialité
- R3 : linguiste anglophone, sans expérience d'enseignement, qui évolue dans le milieu de l'édition
- R4: enseignant-chercheur, linguiste et anglophone, avec expérience tant en recherche qu'en pratique de terrain en anglais de spécialité
- R5 : linguiste, expert en linguistique systémique fonctionnelle

S'ensuit ci-dessous la séquence utilisée pour les tests.

i. Test n°1 (R1, R2, R3) : sans consultation au préalable de la taxonomie des erreurs
ii. Test n°2 (R1, R2, R3) : après signalement explicite d'items erronés (sans les étiquetages)
iii. Test n°3 (R1, R2, R3, R4) : après signalement explicite avec les différents étiquetages
iv. Test n°4 (R1, R2, R3, R4) : identique au test n°2, mais avec les erreurs d'acceptabilité[82]
v. Test n°5 (R1, R2, R3, R4) : identique au test n°3, mais avec les erreurs d'acceptabilité
vi. Test n°6 (R1, R5) : après signalement explicite des tags de la linguistique systémique

Tout d'abord, pour le test n°1, il s'agit d'une étape de reconnaissance dans laquelle il a été demandé aux annotateurs (codeurs) R2 et R3 d'identifier les items jugés erronés, dans 24 des 244 textes qui constituent notre corpus. Cela signifie principalement qu'il fallait indiquer si un élément devait être considéré comme une erreur ou non, mais sans en préciser la typologie d'erreur que l'on aurait attribuée au préalable (mais séparément) à l'item concerné. La deuxième étape, ou le test n°2, était celle portant sur l'accord des items signalés par R1 en tant qu'erreurs. Il était question ici pour les annotateurs d'accepter (ou non) les items signalés. Les deux étapes ont été effectuées pour le volet portant sur les erreurs du système linguistique[83] - avec le schéma d'UAM. Dans l'étape initiale (test n°1) il s'agissait tout simplement donc d'un signalement alors que dans la deuxième partie (test n°2), les annotateurs devaient se prononcer sur le signalement du R1.

La troisième étape ou le test n°3 portait sur l'accord des items pré-étiquetés et signalés en tant qu'erreurs. Il était question ici pour les annotateurs d'accepter (ou non) les items signalés et les annotations correspondantes. La différence notable entre les tests n°2 et n°3 réside dans le fait que les annotateurs pouvaient désormais librement modifier les choix du R1 et ses étiquetages : ce choix permet d'explorer le niveau de concordance entre le fait d'accepter qu'un item soit étiqueté

[82] Les erreurs d'acceptabilité textuelle renvoient aux erreurs non-grammaticales et donc l'identification en tant que telle est intrinsèquement liée à l'environnement textuel immédiat. Voir chapitre VI pour plus de précisions.
[83] Ce qui est signalé ici est à l'opposé des erreurs d'acceptabilité textuelle. Mais soulignons que ces différentes catégorisations reposent principalement sur les étiquetages de l'UAM CorpusTool. Cf. chapitre V pour plus de précisions.

en tant qu'erreur jusqu'à lui attribuer une typologie similaire. Les tests n°4 et n°5 portent uniquement sur le deuxième volet des annotations – à savoir, uniquement sur les erreurs d'acceptabilité textuelle par opposition aux erreurs propres au système linguistique. De plus, les tests n°4 et n°5 reproduisent les procédés du test n°2 et du test n°3 respectivement, tandis que le test n°6 amène l'annotateur (R5) à évaluer les étiquetages issus de la linguistique systémique fonctionnelle et qui ont été préalablement faits par nos soins en tant qu'annotateur (R1).

Toutefois, avant de passer aux détails des différents tests, précisons simplement que le calcul du degré d'accord entre annotateurs a été effectué à l'aide d'une mesure statistique appelée le Kappa de Cohen[84]. L'avantage de cette méthode de calcul est de fournir une estimation d'accord facilement compréhensible par la communauté scientifique plus large et qui, de plus, est jugée incontestable – ce qui n'est pas tout à fait le cas avec les accords inter-annotateurs fournis en pourcentage. En effet, Le Kappa de Cohen permet de mesurer l'accord entre deux annotateurs en calculant aussi bien (i) la proportion de l'accord observé dit aussi l'accord relatif (AO/P*a*) et la probabilité que l'accord soit aléatoire (AA/P*e*).

Le tableau ci-dessous de Landis & Koch (1977) illustre à titre d'information une interprétation possible des scores ou coefficients obtenus par les deux mesures statistiques. Notons cependant qu'il n'y a pas de consensus absolu sur la gamme de valeurs proposées et que l'interprétation peut varier dans la littérature de quelques points suivant les auteurs.

Score de Kappa	Degré de concordance
< 0.00	médiocre
0.00 — 0.20	minime
0.21 — 0.40	juste
0.41 — 0.60	Modéré
0.61 — 0.80	Considérable, solide
0.81 — 1.00	Presque parfait

Tableau 12 : l'échelle des coefficients de Kappa

4.3.2 Est-ce bien une erreur ? Quelle concordance entre annotateurs ?

Cette section passe en revue les résultats des tests n°1 et n°2 et implique de ce fait R1, R2 et R3. A titre d'illustration, deux éléments sont fournis dans le but de mieux expliquer comment nous sommes parvenus aux résultats finaux : (i) tout d'abord une table de contingence appelée aussi

[84] Le choix de la mesure s'est imposé pour une raison purement pratique : le kappa de Cohen est la mesure la plus communément employée dans la littérature portant sur l'accord inter-annotateur. De ce fait, les autres mesures telles que le « Scott's Pi » et le kappa de Fleiss qui permettent aussi de calculer l'accord inter-annotateurs ont été écartées. Et ce, étant donné qu'elles sont employées dans une moindre mesure dans la littérature actuelle.

matrice de confusion illustre ci-dessous le nombre d'accord et désaccord enregistré entre R1 et R2 tandis que (ii) le calcul du premier kappa de Cohen a été effectué à partir du premier échantillon numéroté *txt_010_sm1* sur lequel R2 a dû identifier ce qu'il considère comme erreur.

txt_010_sm1			
	R1		
R3	Accord	désaccord	total
Accord	22	11	33
Désaccord	6	293	299
total	28	304	332

Tableau 13 : Exemple d'une matrice à confusion utilisée pour calculer le Kappa de Cohen

Pour rappel, R1 renvoie aux annotations effectuées par nos soins. De ce tableau, il faut donc comprendre que l'échantillon comporte 332 items lexicaux parmi lesquels 293 n'ont pas été signalés comme étant une erreur : ni par R1, ni par R3. Le nombre d'éléments annotés conjointement comme erreur par R1 et R2 s'élève à 22, tandis que R1 considère 6 éléments de plus comme étant erronés contrairement à R2, et ainsi de suite. En utilisant l'équation 1 ci-dessous avec les données du tableau ci-dessus on obtient notre premier score de Kappa (*k).*

$$k = \frac{\sum(Pa) - \sum(Pe)}{n - \sum(Pe)}$$

Équation 1 : Formule de calcul du Kappa de Cohen

En calculant donc le kappa de Cohen, il s'avère que le taux d'accord observé (AO ou P*a*) est de 0,948 et le taux d'accord aléatoire (AA ou P*e*) s'élève à 0,833, ce qui fait que le coefficient final de kappa est de 0,693. Si l'on se réfère donc à la grille fournissant l'échelle des coefficients de Kappa de Landis & Koch (1977), l'accord entre R1 et R3 sur l'échantillon *txt_010_sm1* constitue un accord tout à fait 'considérable' ou 'solide'. Cela dit, nous pouvons passer maintenant assez succinctement sur l'ensemble des accords constatés à ce niveau du test. En effet, sur l'ensemble des échantillons qui ont été soumis aux deux premiers annotateurs invités pour les tests n°1 et n°2, le premier annotateur invité (R2) a enregistré un coefficient total de 0,74 sur une échelle, rappelons-le, de 0 à 1. Le coefficient du deuxième annotateur (R3) était de 0,77. Ces scores qui affichent tous les deux des taux d'accords observés de plus de 0,9[85] sur un total de plus de 2000 items (ré)évalués signifient, à notre sens, que le premier volet de notre annotation est tout à fait « correct » et pourrait même être reconduit avec des résultats similaires par un annotateur indépendant.

[85] Si l'on devait traduire l'accord observé (AO) en pourcentage simple, on pourrait dire qu'il y a plus de 90% d'accord entre les trois annotateurs.

4.3.3 L'étiquetage des erreurs : quelle fiabilité entre annotateurs ?

De même, plusieurs tests ont été menés en parallèle pour examiner le niveau d'accord entre les étiquetages choisis pour les erreurs dites du système linguistique (cf. chapitre V) et les erreurs dites d'acceptabilité textuelle (cf. chapitre VI). Ces tests concernent R1, R2, R3 et R4. Les taux d'accords observés sont les suivants : entre R1 et R2 = 0,96 ; R1 et R3 = 0,93 ; R1 et R4 = 0,98. De plus si l'on compare les 5 tests, l'accord général indique une tendance similaire entre les coefficients de Kappa : à savoir R1 et R2 = 0,823 ; R1 et R3 = 0,825 et entre R1 et R4 =0,89. Il en ressort deux tendances non-négligeables de l'ensemble de ces résultats : (i) les étiquetages utilisés ont été à la fois compris et se sont révélés tout à fait appropriés par l'ensemble des annotateurs, ce qui explique le taux d'accord élevé à ce niveau et (ii) nous soutenons par conséquent que les mêmes schémas d'annotations pourront être réappliqués tels quels dans un corpus similaire pour arriver à des résultats comparables.

4.3.4 L'étiquetage issu de la linguistique systémique fonctionnelle est-il fiable ?

Le test n°6 est le dernier des vérifications effectuées et il porte sur les annotations issues de la linguistique systémique fonctionnelle. Pour mener à bien ce test, une série de 10 textes ont été soumis à un expert en linguistique systémique fonctionnelle (signalé comme R5) : de ces 10 textes, 5 ont été préalablement annotés par nos soins selon le schéma expérientiel et les 5 restants selon le schéma textuel. Par conséquent, l'expert devait se prononcer uniquement sur les étiquetages signalés, soit en les acceptant individuellement soit en les rejetant et/ou en les modifiant. Le résultat des vérifications effectuées a été confondu dans une table de contingence, comme celle ci-dessous, à partir de laquelle on a calculé le Kappa de Cohen.

	R1					
R5	single top	multiple top	interpersonal	textual	rheme	*total*
single_top	6	0	0	0	0	*6*
multiple_top	0	0	0	0	2	*2*
interpersonal	0	0	0	0	0	*0*
textual	0	0	0	3	0	*3*
rheme	1	0	0	0	65	*66*
total	*7*	*0*	*0*	*3*	*67*	*77*

AO (P*o*)	0,961039
AA(P*e*)	0,7544274
k	0,8413462

Tableau 14 : Illustration d'un score du kappa pour le test n°6 (textuel)

Sont signalés donc dans ce tableau, à titre d'illustration, les résultats des vérifications portant sur le schéma textuel de trois textes. Les quatre premiers étiquetages ou valeurs – de 'single_top' à 'textual' – renvoient aux différentes sous-catégories de la position de thème et le dernier au rhème. Cela dit, si l'on souhaite identifier par exemple les chiffres significatifs dans la colonne verticale 'single_top', on soulignera que le 6 renvoie au nombre d'items sur lesquels les deux annotateurs ont été d'accord, tandis que le 1 renvoie à un étiquetage en tant que 'thème topical individuel' par l'annotateur R1 et en tant que 'rhème' par l'annotateur R5.

Notons que de manière à éviter d'avoir une table de contingence à rallonge, il a été décidé de ne retenir pour le calcul que les couches allant d'une profondeur de 1 à 3 : c'est-à-dire les grandes catégories et non pas l'ensemble des sous-catégories individuelles qui les composent. En effet, aller au-delà de cette profondeur aurait permis d'améliorer l'exactitude du score de Kappa mais celui-ci n'aurait pas été très différent du score obtenu avec une « granularité plus fine ». Autrement dit, avoir une profondeur de 4 ou de 5 aurait permis de voir directement dans la table de contingence ou la matrice de confusion (cf. tableau 14) le type d'étiquetage exact qui avait été attribué par R1 et R5, pour les 3 items signalés comme 'thèmes textuels'. Cela étant, nous avons retenu le même procédé pour le calcul et la visualisation de l'ensemble des tests effectués par R5, aussi bien pour le schéma textuel que le schéma expérientiel. Un exemple est également fourni pour illustrer le calcul sur ce dernier schéma.

	R1			
R5	Process	Participant	Circumstance	total
Process	20	0	0	20
Participant	0	34	0	34
Circumstance	0	3	18	21
total	20	37	18	75
AO (P*o*)	0,96			
AA(P*e*)	0,361956			
k	0,937308			

Tableau 15 : Illustration d'un score du kappa pour le test n°6 (expérientiel)

En ne prenant donc que les grandes catégories, on obtient un score de kappa qui s'élève à 0,937 que l'on pourrait arrondir à 0,94. Cela signifie un accord très élevé, puisque non seulement l'accord observé est haut mais l'accord par chance ou l'accord dit aléatoire est très faible. Toutefois, il convient de signaler que hormis les désaccords signalés par le chiffre 3 dans la colonne verticale 'participant', d'autres désaccords sont également à signaler au niveau du chiffre 34 – correspondant à des accords généraux pour des items signalés en tant que participants. La table de contingence ci-

après avec une granularité plus fine fait office de zoom sur les désaccords précis observés sur l'annotation des participants.

	R1							
R5	material	behavioural	mental	verbal	relational	existential	other	*total*
material	11	0	0	0	1	0	0	*12*
behavioural	0	0	0	0	0	0	0	*0*
mental	0	0	3	0	0	0	0	*3*
verbal	0	0	0	4	0	0	0	*4*
relational	0	0	0	0	12	0	0	*12*
existential	0	0	0	0	0	0	0	*0*
other	2	0	0	0	1	0	0	*3*
total	*13*	*0*	*3*	*4*	*14*	*0*	*0*	*34*

AO 0,88235294
AA 0,30190311
k 0,8314746

Tableau 16 : Précisions sur le score de Kappa (expérientiel : participant)

Même si l'on remarque un écart dans le score de kappa dans les deux précédentes tables de contingence, l'accord global de l'ensemble des vérifications est supérieur à 0.8 ce qui en fait un accord très élevé selon la grille d'interprétation de Landis & Koch (1977). On pourrait donc conclure que l'étiquetage proprement systémique demeure compréhensible voire très accessible aux annotateurs familiers des cadres de la linguistique systémique fonctionnelle, malgré les quelques écarts signalés dans les vérifications de R5. Soulignons toutefois notre réserve quant à l'applicabilité voire la compréhension des étiquetages systémiques par un public non initié à ce courant linguistique.

4.3.5 Le bilan de l'ensemble des tests d'accord inter-annotateurs

Ces six tests de validité, qui ont fait intervenir quatre annotateurs supplémentaires, avaient des objectifs multiples : (i) voir si les mêmes items sont signalés en tant qu'erreur par des anglophones indépendamment de leurs profils et bagages linguistiques respectifs (ii) voir à quel point la typologie d'erreurs attribuée à ces items erronés peut varier d'un annotateur à un autre (iii) et surtout voir si l'idée qu'on se fait d'un étiquetage, c'est-à-dire sa valeur profonde, est suffisamment claire pour être réemployée par une tierce personne face aux mêmes items. Intuitivement on aurait eu tendance à croire que plus les étiquetages sont nombreux, voire plus ils sont pointilleux ou spécialisés, moins il y aurait d'accord. Nos tests d'accord inter-annotateurs ont démontré que cela n'est pas le cas. Pour rappel l'accord général entre R1 et R2 et R1 et R3, avec plus de 2000 items

(re)vérifiés, s'élève à 0,823 et 0,825 respectivement. L'accord entre R1 et R4 est de 0,896 tandis qu'entre R1 et R5 il est de 0,841. Tout cela traduit une très solide validité entre les annotations effectuées par nos soins en tant que R1. De plus, ces résultats signifient par extrapolation qu'un autre chercheur pourrait reproduire notre étude et obtenir des étiquetages quasi-identiques en nombre et en genre à ceux que l'on va décrire dans les chapitres V et VI.

(Chapitre V) Résultats des erreurs du système linguistique

Ce chapitre synthétise les principaux résultats obtenus dans le premier volet d'annotation de notre étude. Dans ce volet, il s'agit principalement de repérer et étiqueter les éléments comme étant erronés à la fois *en* et *hors* contexte. A ce titre, la section 5.1 introduit les étiquetages obtenus par le biais du schéma d'erreur intégré à l'UAM CorpusTool. Les différents signalements d'erreurs dans cette section correspondent de ce fait aux étiquetages propres au schéma et ils sont présentés principalement au moyen de statistiques descriptives. En effet, cette étape renvoie dans une certaine mesure à la première des deux étapes qui constitue les méthodes d'analyses d'erreur selon Granger & Monfort (1994) : à savoir une étape essentiellement descriptive suivie d'une étape explicative. Soulignons, en outre, qu'il est question de présenter ici de façon analytique ce que l'on nomme les erreurs du système linguistique. C'est-à-dire les écarts observés par rapport à une certaine norme linguistique.

Les sections 5.2 et 5.3 retracent, quant à elles, les annotations issues de la section 5.1 qui ont été de nouveau annotées avec une deuxième et troisième couche d'annotation. Ces couches supplémentaires correspondent aux deux schémas inspirés de la métafonction expérientielle et la métafonction textuelle : 5.2 dresse tout particulièrement un état des lieux des problèmes liés à la transitivité tandis que 5.3 ré-explore les mêmes résultats qui sont, à leur tour, observés et signalés en fonction de leur agencement textuel dans la phrase, notamment en position de *thème*. Ces ré-annotations fournissent un aperçu d'une façon novatrice de classifier et d'interpréter les erreurs ; et ce, en utilisant non seulement la syntaxe comme pierre angulaire de la classification mais les différentes métafonctions des items étiquetés.

5.1 Les schémas d'annotation[86] d'UAM CorpusTool v.2.8

Il conviendrait tout d'abord de rappeler que les erreurs qui nous intéressent ici ont été identifiées puis annotées en utilisant le schéma[87] d'annotation présenté en section 4.2.4.1 et illustré par la figure 21 (cf. Annexe : A3 pour le schéma complet). Par ailleurs, nous attirons l'attention dans un premier temps sur les résultats obtenus, non seulement par les annotations ou étiquetages qui traduisent la profondeur des différents niveaux du schéma employé, mais plutôt sur « les têtes de liste » de nos six catégories principales ; à savoir les erreurs lexicales, grammaticales, pragmatiques, de mise en phrase, de ponctuation et celles qui s'avèrent 'incodables'. Il est également à préciser que les erreurs ne seront pas hiérarchisées selon l'importance que l'on y accorde mais tout simplement selon l'ordre donné sur le schéma d'annotation du départ. Cela étant, en termes d'erreurs comptabilisées en semestre 1 (sm1) et semestre 2 (sm2), nous obtenons les tableaux suivants qui mettent en avant les différences observées sous forme de statistiques brutes.

Types d'erreurs	**Sm1**	**Sm2**	**Ecart en *nb***	**Ecart en (%)**
Er. lexicale	560	447	113	20,2%
Er. grammaticale	1589	1518	71	4,5%
Er. pragmatique[88]	226	159	67	29,6%
Er. phraséologique[3]	393	294	99	25,2%
Er. incodable	14	8	6)	42,8%
Er. de ponctuation	124	78	46	37,1%
total	*2906*	*2504*	*402*	*13,8%*

Tableau 17 : Répartition des erreurs du système linguistique comptabilisées avec l'UAM CorpusTool (volet 1)

Toutefois, avant de nous attarder sur le sens réel de ces résultats, il nous paraît judicieux d'expliquer comment nous avons procédé au calcul mais surtout comment nous sommes arrivés à l'interprétation générale qui s'ensuit. En effet, si l'on considère que les participants de notre étude sont suivis de manière semi-longitudinale – c'est-à-dire à travers leurs productions qui sont écrites en anglais langue étrangère et qui sont recueillies à des intervalles différents – les résultats de sm1 et de sm2 peuvent être comparés à des groupes ou échantillons appariés. Cela signifie tout simplement que l'on a des résultats d'un même participant à deux moments différents. Par conséquent, au vu des chiffres obtenus, un *test de student* a été effectué dans le but de réfuter (ou

[86] Dans un souci de clarté, les noms utilisés sur le schéma de départ sont en anglais et sont retenus tel quels notamment dans les tableaux et figures. Ils seront repris, traduits ou explicités dans le corps du texte.
[87] Cf. Annexe (A3) pour voir l'ensemble des niveaux dudit schéma.
[88] Étant donné les chevauchements observés entre ce qui relève de l'ordre du grammatical et du textuel, ces erreurs sont présentées dans le chapitre VI (cf. sections 6.1.1 et 6.1.2) qui traite les erreurs principalement textuelles.

non) l'hypothèse nulle selon laquelle il n'y aurait pas d'écart significatif entre les deux semestres et que la distribution des erreurs entre les deux semestres serait de ce fait normale.

En calculant donc au préalable les moyennes des différentes catégories, telles qu'affichées dans le tableau 17 précédent, nous avons obtenu les premiers calculs nécessaires pour effectuer le *test de student*. Une fois la formule ci-après appliquée, nous arrivons au résultat que t = 3,83 ce qui est illustré dans la figure 25 ci-après.

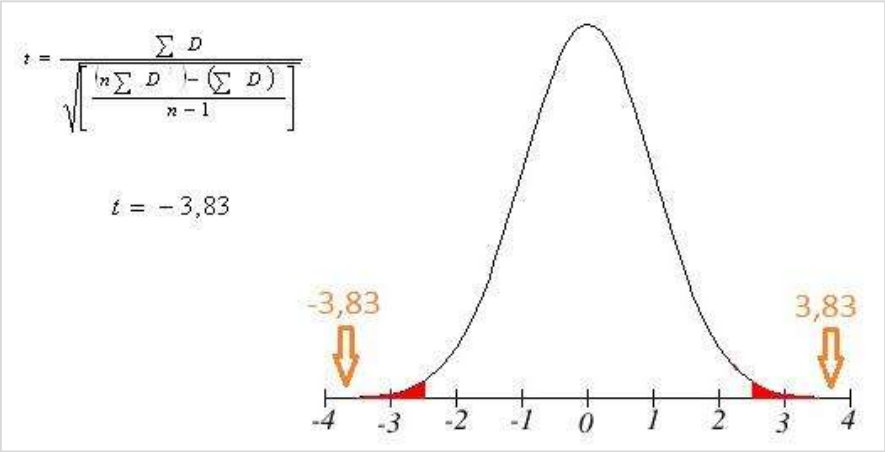


Figure 25 : Calcul et résultat du *test de student*

En comparant donc ce résultat à la valeur critique[89] à ne pas dépasser si l'on veut réfuter l'hypothèse nulle, il devient clair que l'écart observé entre nos deux semestres (i) est bien réel et (ii) est porteur de sens. Cela étant, la valeur de *t* illustrée dans la figure ci-dessus se situe en dehors du seuil critique (de 2,571 si l'on veut avoir un degré de confiance à 95%) et revêt un caractère très significatif, en ce qu'elle indique que la distribution (des erreurs) n'est pas normale et par conséquent que les différences observées entre les deux semestres ne sont pas aléatoires. Notons par ailleurs que la valeur de *p* est de *0,0121* et que l'écart est, de ce fait, compris comme étant indubitablement statistiquement significatif, ce qui laisse entendre qu'il y a une nette progression du sm1 au sm2 chez les sujets-participants.

Toutefois, pour revenir au tableau 17 qui récapitule l'ensemble des résultats de sm1 et sm2, d'autres remarques s'imposent. Par exemple, en observant les quatre premières catégories du tableau - qui sont ordinairement les plus utilisées dans les projets d'analyses d'erreurs - on peut relever une diminution nette au deuxième semestre d'environ 19% en moyenne. Cependant la catégorie où l'on s'attendait intuitivement à observer une réduction plus marquée est celle où l'inverse s'est produit : à

[89] Cette valeur a été obtenue en utilisant un *test de student*, et un test t bilatéral a été effectué avec les paramètres suivants : n=6 ; le degré de liberté = (n-1=5) ; pour un intervalle de confiance à 95% (*p < 0.05)

savoir dans la catégorie grammaticale où la régression s'élève à 4,5%. On s'attardera dans les sections qui suivent à décrire l'ensemble des sous-catégories qui regroupent les six grandes classifications d'erreurs et notamment à mettre en exergue les différents écarts observés.

5.1.1 Les erreurs lexicales

En comparant le nombre total d'erreurs lexicales comptabilisées dans les deux semestres, on observe une diminution nette de 20,17% en sm2 par rapport au sm1. La répartition de ces erreurs se trouve dans le tableau 18. Celui-ci met en avant, entre autres, le fait que la réduction observée n'est pas homogène dans l'ensemble des sous-catégories et qu'il y a même une catégorie (à savoir au niveau des erreurs portant sur le choix lexical – dit aussi de vocabulaire) où l'on constate une tendance inverse.

Catégories d'erreurs lexicales	**sm1**	**sm2**	**D**	**%**	
spelling	353	301	-52	14,7	↘
False-friend	22	3	-19	86,4	↘
coinage	31	22	-8	25,8	↘
borrowing	20	14	-7	35	↘
other-word-choice	86	52	-34	39,5	↘
vocab-choice	48	55	7	14,6	↗

Tableau 18 : Répartition des erreurs lexicales

Parmi les écarts observés entre les deux semestres, deux remarques s'imposent vis-à-vis des erreurs d'orthographe et celles portant sur le choix lexical (ou le choix du vocabulaire). La première de ces catégories enregistre un total de 654 occurrences, ce qui en fait un problème non-négligeable pour nos sujets-participants. La deuxième catégorie, avec beaucoup moins d'erreurs, compte tout de même 103 items auxquels on pourrait ajouter les « faux-amis », « calque » et « emprunt » qui s'apparentent tous – dans une certaine mesure – à des erreurs de choix lexical. Donc, nous avons la possibilité ici de catégoriser l'ensemble des erreurs dans deux grands groupes : à savoir l'orthographe d'un coté, et toutes les autres catégories de l'autre.

Récurrence d'un même item	**Sm1**	**Sm2**
≥ x6	4	2
x5	2	1
x4	3	2
x3	4	5
x2	26	24
x1	232	211
nombre d'items différents	*271*	*245*
Total d'items relevés	*353*	*301*

Tableau 19 : La fréquence individuelle des erreurs lexicales

De ces deux premiers constats, la première chose que l'on remarque quand on s'intéresse de près aux erreurs d'orthographe, c'est la non-systématicité des items relevés. Autrement dit, au niveau des occurrences individuelles annotées, la majorité relève des erreurs que l'on pourrait appeler « solitaires », se produisant une ou deux fois de manière identique (comme indiqué dans le tableau 19). On remarquera par exemple en sm1 la présence de 232 items lexicaux qui ne se répètent pas du tout dans l'ensemble du corpus. Tandis que le nombre d'items lexicaux qui se répètent plus de 5 ou 6 fois demeure également peu élevé. Ce constat conforte la thèse d'une non-systématicité et donc d'une erreur purement humaine par opposition à une erreur d'apprentissage ou une erreur due à la non-maîtrise d'une règle linguistique donnée : le trait distinctif ici pour ainsi dire est donc l'aspect hautement aléatoire de ces erreurs. Nous allons maintenant porter notre attention sur la composition détaillée de l'ensemble des erreurs étiquetées 'lexicales'.

5.1.1.1 « Spelling errors »

Il est à noter ici que 232 sur les 353 erreurs (cf. tableau 19 ci-dessus) signalées en sm1 sont des erreurs qui se sont produites une seule fois ; tandis que sur ces mêmes 353 erreurs, 271 sont des items lexicaux différents. L'écart de 82 (à savoir 23% de l'ensemble) renvoie aux items qui se sont distingués par leur récurrence. C'est-à-dire, à titre d'exemple, qu'en sm1 quatre items se sont produits avec une valeur ≥ 6x : à savoir à une fréquence égale ou supérieure à six fois. Ces quatre items s'expliquent de la manière suivante : *developping* comptabilisé six fois ; *developped* sept fois ; *interdependance* comptabilisé huit fois ; et *interdependant* quatorze fois. En bas, donc, de cette échelle de fréquence se trouvent des erreurs lexicales ayant une récurrence nulle (x1). Quelques exemples sont donnés ci-après.

compagny	*paradoxaly*	*worste*
caracteristics	*agrements*	*comit*
desappearance	*middel*	*Theorically*
Polution	*compains*	*wich*
sovereignity	*desastrous*	*bureaucraty*

Tableau 20 : Quelques exemples d'erreurs d'orthographe

En définitive, nous soulignons que les erreurs d'orthographe se sont réduites de manière globale d'environ 15%. De surcroit, étant donné la non-systématicité des erreurs observées ici, une description approfondie sur la caractérisation de cette diminution s'avère problématique, en raison du fait que seulement deux sur les soixante-et-un textes annotés en sm1 pour des erreurs lexicales n'en avaient pas d'une part, tandis que seulement cinq sur soixante et un au deuxième semestre n'en avaient pas. On peut donc conclure que cette réduction est largement généralisée, au sens que

la distribution est plutôt uniforme entre sujets-participants et traduit peut-être une faible amélioration dans l'attention accordée à la relecture et l'autocorrection dans les écrits chez les sujets-participants.

5.1.1.2 « False-friend errors »

Pour ce qui est des faux-amis, les chiffres sont beaucoup moins révélateurs. On ne dénombre que 25 cas annotés, dont 22 au premier semestre et 3 au deuxième. De plus, la diminution n'est pas aussi remarquable que pourrait laisser entendre l'écart de 86%. En effet, les 22 cas en sm1 se limitent à seulement 16 participants et à quatre items différents : à savoir 'balance' (x19) ; 'sage' (x1) ; privations (x1) ; 'hazard' (x1). Tandis que les trois cas en sm2 relèvent d'un même sujet-participant, avec les items suivants : 'relativise', 'definitively', 'education'. Ces trois termes ont été utilisés respectivement dans le sens '*to put something into perspective'*, '*once and for all*', et '*upbringing'*. Le problème des faux-amis nous parait donc ici superficiel notamment en raison du fait que ce phénomène renvoie, rappelons-le, aux items lexicaux existant dans au moins deux langues avec une orthographe analogue, mais un sens tout à fait différent. Cela étant, ce type de phénomène ne nécessite pas à nos yeux de faire l'objet d'un cours à part entière mais pourrait tout simplement s'introduire sous forme d'un rappel ponctuel non obligatoire.

5.1.1.3 « Coinage errors »

Les erreurs de 'coinage', rappelons-le, renvoient ici à des emprunts de termes existant en langue française adaptés à la morphologie de la langue anglaise. On comptabilise un total de 52 erreurs de ce type dont 30 au premier semestre et 22 au deuxième, avec un écart s'élevant à 26%. Toutefois, comme ce fut le cas avec les erreurs dit de faux-amis, les 30 erreurs du sm1 ont été repérées chez 17 participants de même que les 22 restants du sm2. Cela traduit un pourcentage de 27,8% - correspondant aux 34 textes sur 122 dans lesquels cette erreur a été signalée. La liste de ces néologismes est fournie ci-après :

altern	*exempl,*	*norvegia,*	*sensibilize,*
benefic	*explication,*	*provocate,*	*solidarian,*
comparate	*foundators,*	*provocated,*	*subordonised,*
considerated,	*futur,*	*provocs,*	*to acceed,*
constat,	*implicts,*	*provoque,*	*to applicate,*
denunciate,	*inegalities,*	*provoqued,*	*to combinate,*
discuted,	*inegality,*	*relativised,*	*to instaure,*
effets,	inoved,	reputate,	traduces,

egal,	*limitate,*	*scarify,*	*underminated,*
evoluate,	*megaconcurrential,*	*sensibilize,*	*unlivinable*

Tableau 21 : Quelques exemples de "coinage"

Ce qu'il faut retenir de cette liste est que les interférences (cf. section 7.2.2) peuvent porter sur différentes classes grammaticales, mais avec une préférence marquée pour le verbe. De plus, certains lemmes sont plus récurrents et donc plus problématiques que d'autres, par exemple le verbe 'provoke' en anglais qui apparait sous trois formes différentes avec une fréquence de 10 sur 52 items observés. Par opposition au point précédent de faux-amis, le phénomène ici – de par sa fréquence – mérite que l'on s'y intéresse en cours de langue puisqu'en généralisant les résultats de nos 122 participants à l'ensemble d'une population d'étudiants d'une même année dans un même établissement, la fréquence de ce type d'erreur pourrait s'avérait conséquente.

5.1.1.4 « Borrowing errors »

Quant aux erreurs d'emprunt, ou, dans une certaine mesure, aux erreurs de calque, entendu ici comme l'emprunt d'un élément existant en langue française mais utilisé sans modification en langue anglaise - on en dénombre un total de 34 : dont 20 en sm1 et 14 en sm2. Les 20 occurrences erronées observées au premier semestre sont réparties entre 15 participants et les 14 du deuxième semestre entre 9 participants.

agricol,	*fondation*	*polluant*
changements	*humains*	*poste*
commun,	*inconvenients*	*processus*
creches	*investissement*	*remplace*
crise	*methodes*	*soumission*
dommages	*ministre*	*survie*
et	*objectif*	*ue*
exemple	*permet*	*xxe*
face to	*phenomene*	

Tableau 22 : Quelques exemples d'emprunt

Il est judicieux de préciser ici que la frontière entre ce type d'erreurs et les deux derniers (à savoir les erreurs dites de *faux-amis* ou de *coinage)* peut paraitre minime. Et ce, étant donné que les calques et faux-amis traduisent d'une manière générale (i) une méconnaissance de l'équivalent lexical d'un terme utilisé dans la langue cible ou (ii) une sorte de piège facile dans lequel tout apprenant inattentif pourrait tomber.

5.1.1.5 « Other word choice errors »

'Other-word-choice' renvoie à ce que l'on désigne communément comme lapsus. Toutefois, il n'y a aucun moyen de vérifier que l'élément erroné relève d'un item lexical non-maîtrisé sémantiquement plutôt que d'un emploi ponctuel par inadvertance, sauf bien entendu à contacter le participant en question afin d'avoir des éclaircissements factuels. De ce fait, comme nous l'avons vu pour d'autres types d'erreurs lexicales, la primauté de ces étourderies nous semble appariée avec le principe de la non-systématicité. Et ce, étant donné la récurrence quasi-nulle observée (x1) pour les items relevés à ce niveau : à savoir une seule fois par type et situation phrastique donnée, indépendamment des participants.

A titre d'exemple, nous avons recensé des erreurs sur l'item lexical 'is'. Dans la première phrase, au vu du contexte informatif et textuel il s'agit d'un lapsus manifeste qui aurait dû être 'if it', tandis que dans la deuxième phrase 'is' doit être remplacé par la préposition 'in'.

1. **It is* $If is$ doesn't save Greece, all the countries will be impacted and it will be [...] (txt_53_sm1)
2. Since the economic and financial crisis **is* in 1929, the western countries [...] (txt_48_sm1)

Le même principe est identique avec les deux exemples suivants portant sur le mot 'they'. Dans les deux cas, il s'agit clairement d'une erreur d'inattention – le premier renvoyait à 'the' et le deuxième à 'there'.

3. Consumers would not even know [about] **they* the possibility for them to buy modern objects. (txt_13_sm1)
4. Nowadays, advertising is almost ubiquitous: when we watch television, **they* $there$ are commercial air times every 15 minutes [...] (txt_29_sm1)

Nous pensons que ce type d'erreurs relève de ce que l'on ne peut ni anticiper ni corriger de façon systématique, en raison du fait qu'il caractérise l'erreur humaine par excellence. Nous signalons tout de même une réduction de 39% par rapport au nombre total d'erreurs observées ici, passant de 86 au premier semestre à 52 au deuxième semestre. Nous supposons par conséquent que cela peut s'expliquer par des raisons purement pratiques : notamment que la diminution d'erreurs ici témoigne du niveau d'attention des élèves et tout singulièrement des stratégies de relecture qu'ils mettent en place pour leurs examens de fin d'année.

5.1.1.6 « Vocabulary errors »

Les erreurs de vocabulaire[90] renvoient aux items lexicaux qui ne sont pas maîtrisés sur le plan sémantique, résultant de ce fait dans ce que l'on appelle un faux-sens ou un non-sens. Autrement dit, le mot est connu par l'apprenant mais son contexte d'utilisation précis pose problème. Le nombre total de ce type d'erreurs s'élève à 103 dont 48 en sm1 répartis entre 30 participants et 55 répartis entre 26 participants en sm2. Notons à titre accessoire que ce nombre total se voit multiplié par neuf si l'on y ajoute les erreurs de choix inappropriés signalés et classés dans le schéma d'UAM uniquement selon la classe grammaticale. Nous reviendrons sur ce « problème » dans la section 8.3.1.1. Mais pour l'instant, examinons ces trois exemples d'erreurs de vocabulaire, telles que prévu par le schéma d'UAM.

5. Some **feminisms* $feminists$ think that fighting for women is a long *processus [sic]*, where people have to *make evolve[sic]* their **consciousness* $change-their-mindset$ and point of view of the society. (txt_042_sm2)
6. Besides, the **paper* $money?$ *of[sic]* multinationals is crucial in the world economy because they hire millions of people in the world and they show in a spectacular way how the international system is globalised. (txt_045_sm1)
7. [...] they would be more emotive [...] and would balance work and family very **hardly* $with difficulty$ (txt_031_sm2).

Le principe de la non-systématicité des items lexicaux s'applique de nouveau dans ce type d'erreurs. Et ce, étant donné que la récurrence d'un même item lexical chez plusieurs participants n'est relevée qu'à une fréquence infime. Toutefois, par rapport aux erreurs aléatoires proprement orthographiques, les éléments relevés ici traduisent également un véritable manque et une faiblesse dans le processus d'apprentissage chez les apprenants. En effet, une erreur orthographique suppose de manière générale que le contenu sémantique est maîtrisé, tandis qu'une erreur de vocabulaire laisse entendre le contraire. De plus, dans l'apprentissage d'une langue étrangère, le lexique constitue de manière générale un poids considérable face à l'orthographe.

5.1.2 Les erreurs grammaticales

Pour les erreurs relevant d'une non-maîtrise ou d'un écart d'usage du système linguistique, la comparaison entre semestres n'a pas été de nature – que nous jugeons – statistiquement

[90] Les problèmes de vocabulaire recensés ne tiennent pas compte de ceux identifiés dans les catégories grammaticales où le problème se pose à la fois au niveau du sens et de la fonction de l'item dans la phrase.

significative : notamment en termes de ce qui avait été escompté. En effet, l'ensemble des erreurs de la catégorie grammaticale n'a diminué que de 4,5% : passant de 1589 items annotés au premier semestre à 1518 au deuxième. Il faut cependant noter que la distribution de cette réduction n'est pas homogène, comme nous le démontrons ci-après.

Catégories d'erreurs grammaticales	**sm1**	**sm2**	***n= (sm1+sm2)***	***d= (sm2-sm1)***	***(%)***
np-error	576	543	*1119*	*-33*	*-5,72*
adjectival-phrase-error	66	50	*116*	*-16*	*-24,24*
adverb-phrase-error	33	22	*55*	*-11*	*-33,33*
prep-phrase-error	185	170	*355*	*-15*	*-8,10*
vp-error	364	395	*759*	*+31*	*+8,51*
clause-error	184	166	*350*	*-18*	*-9,78*
clause-complex-error	164	144	*308*	*-20*	*-12,19*
special-structure-error	12	24	*36*	*+12*	*+100*
other-grammatical-error	5	4	*9*	*-1*	*-20*

Tableau 23 : Répartition des erreurs grammaticales

Le tableau 23 met en avant des données brutes renvoyant au nombre total d'annotations obtenues au premier semestre (sm1) ; au deuxième semestre (sm2) ; au total comptabilisé des deux semestres (n) ; la différence observée entre semestres (d) ; et le pourcentage effectif de la différence observée. Au vu de ce tableau, les points de grammaire qui illustrent les meilleures "améliorations" en pourcentage sont les 'adverb-phrase-error' (33%), les 'adjectival-phrase-error' (24%) et les 'clause-complex-error' (12%): tandis que "les meilleures améliorations" en chiffres bruts sont les "np-error" (-33), les 'clause-complex-error' (-20) et les 'clause-error' (-18). De manière générale, ces réductions ne nous paraissent pas suffisamment significatives, compte tenu de (i) la variable temporelle qui représente un écart d'environ 6 mois entre les rédactions en sm1 et sm2 et (ii) les 2 heures 30 minimum de cours intensif de langue anglaise, suivi de façon hebdomadaire.

Toutefois considérer les éléments ayant les plus hauts pourcentages comme synonymes d'amélioration peut induire en erreur, en raison du fait que ces pourcentages reflètent des catégories avec un faible nombre de cas annotés en leur sein comparé aux autres sous-catégories. Autrement dit, les 33% de réduction observés dans la catégorie 'adverb-phrase-error' en sm2 doivent être relativisés au vu du bien maigre nombre d'items annotés au premier semestre. Ceci n'est pas en soi une mauvaise chose. La figure 26 ci-dessous permet de voir ces réductions de façon proportionnelle : à savoir le total obtenu en sm1 et la différence observée entre les deux semestres.

Au travers de ces visualisations, il devient clair que la baisse observée est très relative voire minime dans l'ensemble des différentes sous-catégories d'erreurs grammaticales. En effet, l'écart maximum observé dans l'ensemble des sous-groupes est de 33 (dans le groupe 'np-error') alors que le nombre total d'occurrences erronées enregistrées dans la catégorie correspondante en sm1 était de 576. Ce qui en fait une baisse de 5%. De plus, le faible écart global va à l'encontre de ce qui avait été escompté, nous obligeant en conséquence à nous poser davantage de questions sur ces résultats statistiques. Nous allons donc, dans un ordre décroissant, détailler l'ensemble de ces différentes catégories de 'np-error' en fonction du nombre total d'items comptabilisés par sous-catégorie.

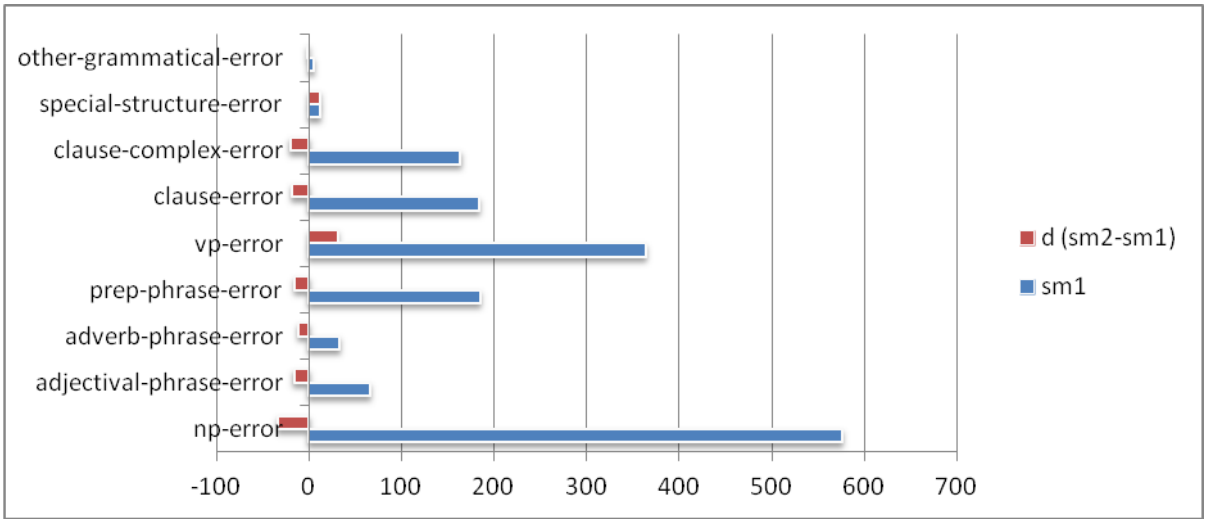


Figure 26 : L'écart relatif des erreurs grammaticales entre les semestres

5.1.2.1 « Np-error »

Tout d'abord, rappelons que la catégorie d'erreurs représentée ici est celle du groupe ou syntagme nominal. Elle porte principalement sur le noyau du groupe nominal – c'est-à-dire 'la tête'– avec ses rapports privilégiés avec des satellites à gauche (les déterminants et les pré-modifieurs) et à droite (les post-modifieurs). Notons également que dans cette catégorisation, il y a des erreurs de pronom annotées comme telles quand l'erreur porte sur l'item en fonction de 'tête'[91].

les sous-catégories 'np-error'	**sm1**	**sm2**	***n(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
determiner-error	355	268	623	-87	-24,50
premodifier-error	19	31	50	+12	+63,15
head-error	153	220	373	+67	+43,79
postmodifier-error	22	10	32	-12	-54,54
np-complex-error	1	1	2	0	0

[91] Notons que les erreurs de pronom sont étiquetées selon leur nature et peuvent par exemple être signalées comme erreur de référence par opposition à un simple choix erroné ayant trait à la classe grammaticale employée.

pronoun-error	26	13	39	-13	-50

Tableau 24 : Répartition des erreurs du groupe nominal

Comme nous l'avons expliqué précédemment, la comparaison entre semestres peut facilement induire en erreur si l'on commence l'analyse à partir des pourcentages affichés à droite des tableaux. Il convient de ce fait de commencer avec les colonnes *d(sm2-sm1)* et *n(sm1+sm2)* qui indiquent respectivement la différence effective de manière chiffrée entre nos deux semestres d'observation et le total d'éléments annotés. Ces deux colonnes nous permettent donc d'établir l'ordre décroissant des différentes catégories en fonction de leur « gravité » : c'est-à-dire celles qui mériteraient que l'on y accorde le plus d'importance correctrice dans une classe de langue.

(np-errors → erreurs portant sur les déterminants)

Intéressons-nous de ce fait aux erreurs de déterminants qui, rappelons-le, ont pour fonction de déterminer et limiter la portée du groupe ou syntagme nominal qu'ils précèdent. En effet, à en croire les deux tableaux précédents, les déterminants constituent une sorte de talon d'Achille de nos sujet-participants. Ces éléments épineux constituent l'erreur la plus récurrente en termes statistiques : par exemple elle représente 55,6 % des 1119 erreurs identifiées dans le groupe nominal (ce qui ne constitue pas une surprise en soi au vu de la composition de base du syntagme nominal) ou encore 20% de l'ensemble des 3107 erreurs annotées dans la catégorie grammaticale. Aucun autre élément n'obtient de telles statistiques tout seul dans l'ensemble de notre corpus. Par ailleurs, il est important de souligner ici que la majorité des cas annotés portent sur l'article défini, mais que même dans ce sous-ensemble d'erreurs les caractéristiques des erreurs ne sont pas homogènes.

Cela étant, les quatre types d'erreurs de déterminants les plus fréquents sont précisés et sont suivis d'exemples ci-après :

(i) la présence non-requise d'un déterminant, il s'agit principalement de l'article défini (t*278 occurrences*) : l'erreur dans l'usage survient par exemple (i) lorsque le sujet nominal n'est pas connu de l'allocutaire et n'a pas été mentionné précédemment ; (ii) lors qu'il s'agit de généralités ou d'abstraction ; ou encore (iii) lorsqu'il s'agit de parler de certains pays ou des saisons.

8. Universities […] try to adapt themselves but they can't foresee eventual changes in the world of **the* $~~the~~$ work. (txt_015_sm1)
9. We are *on* the 1st of February, and **the* $~~the~~$ winter is eventually showing up! It is more than one month late. The effects of **the* $~~the~~$ climate change are affecting more and more our daily life. (txt_016_sm1)

10. We can think about the situation of **the* $~~the~~$ Spain [...] (txt_023_sm1)
11. Some solutions *has[sic]* been found like the using of other energy resources. For instance **the* $~~the~~$ wind power (windturbine), **the* $~~the~~$ solar energy (solar panel) or even synfuel. Moreover, **the* $~~the~~$ oil is getting more and more expensive [...] (txt_011_sm1)

(ii) <u>l'absence d'un déterminant requis</u> (*t222 occurrences*). Dans ce contexte, l'article est obligatoire et ne peut en aucun cas faire l'objet d'une ellipse ou ne relève pas d'un usage dit de l'article zéro.

12. With **Ø* the Greek crisis, we can see that the solution has to come from the European Union [...] (txt_039_sm2)
13. But we will see that some criteria or trainings have been developed for **Ø* the last 40 years. (txt_008_sm2)
14. But there is **Ø* an economic sector where supply is growing very quickly because of an absent demand. (txt_035_sm2)
15. Europe was situated between **Ø* the United States and the *communism bloc[sic]* and had to prove that it was *powerfull[sic]*. (txt_029_sm2).

(iii) <u>Le choix erroné de déterminant</u> (*t82 occurrences*) : l'erreur se porte ici sur le choix inapproprié pour plusieurs raisons. 42% des cas recensés portent sur les articles indéfinis : à savoir '*a*', '*an*' et '*any*'. De plus, même ceux portant sur 'a' ne sont pas tous du même type, comme illustré ci-après.

16. Knowing that there is a problem is **a* one thing but knowing how to solve it is *an other[sic]* one. (txt_035_sm1)
17. [...] because of **a* an economic crisis, they can't afford to buy our products [...] (txt_045_sm1)
18. According to **a* an IPPC's survey, people have to reduce *60%[sic]* their greenhouse gas emissions. (txt_055_sm1)
19. The idea of **an* a European Union is really old [...]. (txt_039_sm1)

(iv) <u>L'accord du déterminant</u> [avec le nom] (*t36 occurrences*) : la quasi-majorité des cas rapportés concernent l'usage erroné des démonstratifs, notamment 'this' et 'that'. Le problème le plus souvent relevé est lorsqu'il y a un manque d'accord en nombre avec le noyau du syntagme

nominal au pluriel et un déterminant au singulier. Le cas où ce dernier est au pluriel avec un nom singulier a été très peu relevé.

20. Then, the government issues a credit to **that* $those$ companies *to[sic]* their emissions. (txt_056_sm1)
21. Besides women are expected to do the majority of **this* $these$ household tasks [...] (txt_056_sm2)
22. The employment of women is one of the great changes of **this* $these$ last 50 years. (txt_009_sm2)

Le fait que les déterminants posent problème aux apprenants est un constat connu et partagé par les enseignants d'anglais langue étrangère. Cependant, nous ne nous attendions pas à une telle profusion par rapport aux différentes catégories existantes dans le groupe nominal ou encore moins à voir que les erreurs de déterminants constitueraient un pourcentage si élevé par rapport à la totalité des catégories grammaticales explorées. Force est cependant de constater que ces résultats vont dans le sens de certains ; VanDyke & Lehman 1997 ; Miller 2005 ; Barrett & Chen 2011 soutiennent que l'acquisition du système de déterminant constitue un problème majeur pour tous les apprenants d'anglais langue étrangère. Ce point épineux est d'autant plus problématique quand l'emploi de déterminants diffère dans la langue source par rapport à la langue cible.

(np-errors → erreurs portant sur les n-têtes)

Quand l'erreur porte sur le noyau même du syntagme nominal (appelé ci-après le n-tête), celle-ci a été annotée et caractérisée de quatre façons différentes.

(v) <u>Accord en nombre.</u> Le problème le plus récurrent ici, avec 250 occurrences observées, relève d'une erreur portant sur le nombre. Dans les cas mis en exergue ci-après, le contexte linguistique exige une précision en nombre qui n'a pas été respectée. Cela peut conduire aux problèmes de n-tête ayant une forme erronée aussi bien au singulier qu'au pluriel, mais dans lesquels l'inverse aurait été la seule option convenable sur le plan grammatical.

23. Since 2008, most *of[sic]* **country* $countries$ are going through […]. (txt_012_sm1)
24. However, we can point out another **arguments* $argument$ which can explain [...] (txt_032_sm1)
25. [...] a program in which a company reduces its carbon emissions is one of the **solution* $solutions$. (txt_034_sm1)
26. To conclude, every **countries* $country$ has to set a program in order to reduce emissions of carbon dioxide [...]. (txt_060_sm1)

(vi) Problème de vocabulaire. Le problème est d'ordre purement sémantique, voire lexical. L'erreur porte sur le choix lexico-fonctionnel du noyau du syntagme nominal.

27. Post-graduate **people* $students$ are *not[sic]* longer efficient in the world of work. (txt_006_sm2)
28. [...] they still cannot achieve their **assertations* $goal$ and make it to the top. (txt_009_sm2)
29. That's why women keep going to fight for their rights, to have the same **ways* $privileges$ as men [...] (txt_024_sm2)
30. Indeed, the European Central Bank set the same **politics* $policies$ for Greece and Germany. (txt_058_sm2)

(vii) Problème du choix de classe grammaticale pour le *n-tête*. Ici, le noyau du syntagme nominal n'appartient pas à la classe nominale. Pourtant, l'item joue le rôle de *n-tête,* ce qui entraine l'erreur d'appréciation.

31. This can really lead to a great **improve* $improvement$ of the ecologic situation. (txt_009_sm1)
32. This declaration represented a huge **threaten* $threat$ for its competitors. (txt_031_sm1)
33. Degrees need to be reformed in order to respect the **changing* $changes$ in the world of work. (txt_041_sm1)
34. [...] and if it offers them some better opportunities, **mostly* $most$ are restricted to a certain category of jobs [...] (txt_024_sm2)

(viii) L'absence de noyau n-tête. Ce qui pose problème ici est le fait que le n-tête est absent. De plus, malgré le fait que l'on pourrait facilement 'deviner' l'item manquant, l'effet d'ellipse ne pourrait être invoqué comme résultat désiré, en raison du fait que l'élément n'est souvent pas suffisamment saillant pour avoir recours à de tels procédés.

35. However, we forgot the most important **Ø* $thing$. (txt_017_sm1)
36. But people ten[d] to be more concerned *by*[sic] the eco-friendly **Ø* $practices$ nowadays and they agree with environmentalist [...] (txt_027_sm1)
37. It would be a good thing for every one if they could live in better **Ø* $conditions$ and afford convenience goods [...] (txt_046_sm1)
38. The majority of postgraduates are now women in the OECD **Ø* $countries$. (txt_056_sm2).

(np-errors → erreurs portant sur les pré-modifieurs et les postmodifieurs)

Viennent ensuite les erreurs qui portent sur les « satellites » du noyau nominal, à savoir celles qui le modifient de manière générale en précisant ou limitant sa portée. Pour les pré-modifieurs, les différentes sous-catégorisations sont au nombre de quatre (exemplifiées de *ix* à *xii*), tandis que pour les post-modifieurs eux sont regroupés en deux groupes (*xiii* & *xiv*). Des exemples précis sont fournis pour chaque cas différent.

(ix) Problème d'agencement. L'accent est mis ici sur des erreurs principalement de type adjectival qui se trouvent en position pré-nominale – ce qui est plus ou moins la norme - alors que la contrainte de la structure du groupe verbal obligerait une position post-nominale. Soulignons toutefois que ces cas sont observés principalement autour de syntagmes nominaux fonctionnant comme « objet » ou « complément » d'un verbe.

39. First, what makes **possible* $globalisation possible$ globalization is the increasing capacity […] (txt_025_sm1)
40. To correct this *inadequation[sic]*, government and education might make **attractive* this university course $attractive$ by proposing […] (txt_035_sm2)

(x) Problème de choix et de formation de la fonction 'pré-modifieur'. La question se pose quand il y a une erreur sur la classe grammaticale. C'est-à-dire l'item jouant le rôle du modifieur n'appartient pas à une catégorie grammaticale permettant d'occuper cette fonction. Cela se produit, à titre d'exemple, quand un adverbe ou un nom se trouve substitué à la fonction adjectivale – provoquant, de ce fait, un refus irréfutable de la proposition (au sens de *clause*) dans son ensemble.

41. Each month *this[sic]* is a **dramatically* $dramatic$ European Union's summit […]. (txt_039_sm2)
42. [...] women are seen as gentle and harmless beings, and many **men* $male$ employers are still questioning themselves (txt_019_sm2)

(xi) Problème de choix lexical. Le refus de l'item est provoqué par un choix lexical inapproprié : il s'agit bien d'un item pouvant jouer la fonction désignée du moins en termes de syntaxe, mais qui demeure sémantiquement inexact.

43. Taking all into consideration, the cartoon represents a **main* $major$ issue in our society […]. (txt_052_sm1)
44. […] education is not the **unique* $only$ factor to be take[n] into account concerning the professional field. (txt_042_sm2)

(xii) Problème de pluralisation. Le problème qui se pose ici est celui de la pluralisation des items ayant la fonction de 'pré-modifieur'.

45. Meetings with directors, managers and **others* $other$ heads of companies are also very instructive. (txt_016_sm2)
46. To conclude, we can see that there is obviously a gap between **universities* $university$ degrees and the world of work. (txt_038_sm2)

* *

*

(xiii) Position d'un groupe grammatical inadapté après le n-tête. Il s'agit principalement d'un groupe adjectival qui survient après le groupe nominal, alors que celui-ci est supposé restreindre ou modifier la portée de ce dernier. Cela crée par conséquent des erreurs d'appréciation en termes d'agencement syntaxique.

47. So we need to find other solutions **more efficient* $more efficient solutions$ because […] (txt_010_sm1)
48. […] the EU has today twenty seven member-states and it is considered as the second territory **most powerful* $second most powerful territory$ in the world after the United States of America. (txt_029_sm2)
49. The EU is not a big country **unified* $big unified country$ but *and[sic]* organisation of countries in which all of them have their own s*overeighty[sic]* […]. (txt_029_sm2)

(xiv) Problème du génitif anglo-saxon ('s). Il est question de l'usage d'une phrase prépositionnelle ou d'une structure (N de N) alors que le génitif est requis.

50. Education, and especially universities, is the "one best way" to develop **abilities of people* $people's abilities$ and […]. (txt_030_sm2)

(np-errors → erreurs portant sur les pronoms)

Les dernières sous-catégories d'erreurs apparentées au syntagme nominal s'articulent autour de celles occupant la fonction de pronom. En effet, il s'agit ici (i) d'erreurs de cas grammatical inapproprié, (ii) d'erreurs d'accord en nombre et (iii) d'erreurs de choix lexical. Parmi ces trois sous-catégories, il est important de préciser qu'il est essentiellement question de l'emploi agrammatical d'un pronom à la place d'un autre. A titre d'illustration, en utilisant le cas nominatif à la place d'un accusatif ou d'un génitif – comme dans l'exemple suivant où l'accusatif est utilisé au lieu du pronom réflexif *'themselves'*: *'But the facts talk for *them'* (txt_021_sm2). Cela conduit à un problème de grammaticalité d'une part et à un problème de référence d'autre part. Mais ce

dernier point ne sera abordé de façon approfondie que dans la section ayant trait aux erreurs de cohésion et de manière plus globale à la notion de référence.

En bref, les différentes erreurs observées sur le n-tête et tout particulièrement sur ses satellites se trouvant à sa périphérie gauche et droite nous indiquent à quel point le problème ne peut se réduire à bien choisir un article et un nom. Les différents exemples nous renseignent sur l'ampleur du problème et constituent en soi une prise de position nécessaire pour mieux orienter et préparer l'étude de l'ensemble de ces points grammaticaux épineux en cours de langue. Cela dit, au vu des effectifs observés dans cette catégorie (pour rappel, déterminants (623) ; prémodifieurs (50) ; n-tête (373) ; post-modifieurs (52) et pronom (39)), il est important de pouvoir considérer ces fréquences relatives comme un indicateur d'un problème de fond (ou comme indice de la gravité du problème) qui nécessite un véritable travail en profondeur et non seulement comme un indicateur du temps à attribuer aux différents points dans un cours de grammaire.

5.1.2.2 « Vp-error »

Examinons maintenant ce qui constitue, en termes de fréquence d'occurrences erronées, le deuxième point le plus épineux pour nos étudiants francophones apprenant l'anglais langue étrangère : à savoir les différentes catégorisations appartenant au syntagme verbal. Cependant avant de nous pencher sur ces dernières de manière détaillée, deux graphiques sont fournis ci-après afin d'avoir une vision d'ensemble sur les différentes annotations et les fréquences correspondantes.

les sous-catégories 'vp-error'	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
verb-vocab-error	169	195	364	+26	+15,4
subject-finite-agreement	101	93	194	-8	-7,92
modal-formation-error	16	25	41	+9	+56,3
do-aux-construction-error	0	2	2	+2	####
perfect-formation-error	9	19	28	+10	+111
progressive-formation-error	5	0	5	-5	-100
passive-formation-error	15	12	27	-3	-20
have-to-constuction-error	0	1	1	+1	####
negation-construction-error	14	15	29	+1	+7,14
interrogative-formation-error	15	11	26	-4	-26,7
modal-tense-aspect-selection-error	2	1	3	-1	-50
multiple-verb-formation-error	3	1	4	-2	-66,7
vp-missing-error	9	9	18	0	0
gerund-construction-error	6	10	16	+4	+66,7

Tableau 25 : Répartition des erreurs du groupe verbal

Pour rappel, la colonne à gauche renvoie à l'ensemble des catégories identifiées en tant qu'erreurs appartenant au groupe verbal ; *'n'* renvoie au nombre d'occurrences comptabilisées par catégorie ; *'d'* présente la différence observée de manière chiffrée entre nos deux semestres d'étude (sm1 et sm2) ; et enfin, la colonne à l'extrême droite met en exergue le pourcentage effectif par rapport à la différence observée. Cela étant dit, ce qui est le plus frappant dans le tableau 23 est l'augmentation du nombre total d'erreurs annotées en sm2 (t394) par rapport au sm1 (t364). Cette augmentation de 8% pourrait laisser croire que la situation a empiré. En effet, uniquement six sous-catégories sur quatorze enregistrent une baisse au sm2 par rapport au sm1 ; une catégorie est restée stable ; et les sept restantes ont vu une augmentation d'erreurs par rapport au sm1. Toutefois le deuxième graphique ci-dessous permet de mieux apprécier la tendance observée.

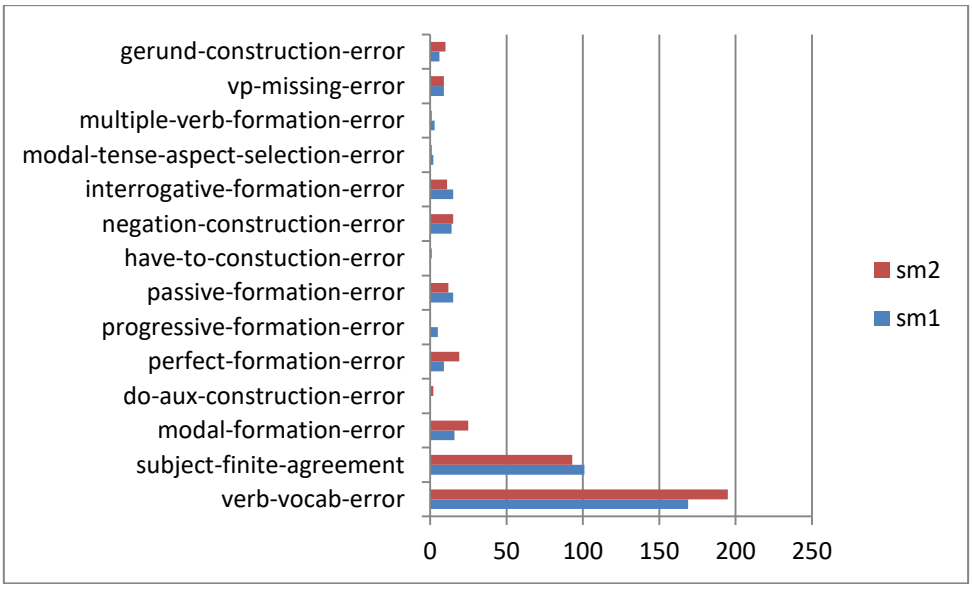


Figure 27: L'écart observé entre sm1 et sm2 dans le groupe verbal

Au vu de cette comparaison, deux catégories sur quatorze enregistrent - à elles seules - 73% des observations relevées ici : il s'agit (i) d'erreurs de vocabulaire portant sur un élément jouant le rôle de 'verbe principal' de la phrase ou (ii) d'erreurs d'accord entre le noyau verbal et le sujet auquel il doit s'accorder. Plus de détails seront fournis ultérieurement sur ces deux cas, avec des exemples correspondants. Quant aux autres erreurs, la figure 27 permet de comprendre que la multiplicité des étiquetages possibles et le faible nombre d'erreurs par catégorie favorisent des biais d'interprétation dans la prise en compte de celles qui se seraient « améliorées » ou « aggravées » entre les semestres. Autrement dit, il serait naïf de vouloir tirer des conclusions sur le 66% à la fois de réductions enregistrées dans la catégorie « multiple-verb-formation-error » ou le même 66% d'augmentation des « gerund-construction-error » - notamment en raison du faible nombre d'observations qui y est retenu par rapport aux autres types signalés dans la même catégorie. Par

conséquent, nous ne nous attarderons que sur les quatre points les plus récurrents, en termes de statistiques brutes.

(vp-errors → erreurs portant sur le choix du verbe lexical)

Notons que les erreurs relevées ici sont différenciées de celles désignées comme étant uniquement « lexicales[92] » en section 6.1.1 et qui - comme nous l'avons expliqué - n'ont donc pas d'incidence sur la grammaticalité de la phrase dans son ensemble.

(i) Le choix erroné du verbe lexical. *(t364 occurrences)* Les erreurs relevées ici sont principalement d'ordre sémantique. C'est-à-dire que la syntaxe de la phrase n'est pas remise en cause. L'erreur porte de ce fait uniquement sur le choix du verbe lexical qui peut provoquer non seulement un non-sens dans la phrase concernée, mais aussi dans celles qui suivent : notamment si le choix lexical concerne un élément clé qui est repris plus tard dans la chaîne informationnelle.

51. Globalisation is **touching* $affecting$ // *all the world[sic]* and won't disappear tomorrow, we have *to *do* $live$ // *it [sic]* with it even it is implies more connection *an[sic]* interdependence in economical spheres. (txt_015_sm1)
52. If the wealthy countries collapse, the countries of the Third-World *are[sic]* concerned too. The households **are involved in* $are affected by$ the crisis. They are first concerned by the austerity package […]. (txt_023_sm1)

La contre-tendance observée ici mérite d'être soulignée dans la mesure où le nombre d'occurrences erronées est plus nombreux au deuxième semestre : à savoir 15% de plus qu'en sm1. Précisons également que l'on comptabilise 190 tokens ou items lexicaux différents sur les 364 observés. Ce qui laisse croire qu'à part un faible nombre de mots récurrents – comme par exemple 'make' (x18) ; 'know' (x16) ; 'do' (x12) ; 'touch' (x12) ; 'follow' (x9) take (x6) – les autres sont beaucoup moins systématiques, voire ne se répètent pas plus d'une fois. Toutefois, la fréquence globale mérite que l'on y accorde de l'importance et par conséquent un temps de correction/rémédiation en cours de langues.

(vp-errors → erreurs d'accord et erreurs diverses)

Comme le titre l'indique, il s'agit tout d'abord de relever les erreurs d'accord observées entre le sujet et le verbe. En ce qui concerne ce dernier, il est question d'erreurs d'accord entre le sujet nominal et l'auxiliaire (comprendre le *finite* ou le *conjugue* selon la terminologie systémique, cf.

[92] Cf. les sections 8.3.1.1 et 8.3.2 au sujet des limites de la classification des erreurs dites lexicales.

section 3.2.5) ou entre le sujet et le verbe lexical. De plus, bien que les erreurs relevées aient été identifiées majoritairement sur des verbes lexicaux individuels ayant une faible récurrence – à savoir entre 1 et 2 fois pour 55% des cas relevés – 45% des cas recensés ici concernent trois verbes précis.

(ii) L'accord entre sujet et auxiliaire. *(t87 occurrences)* Les 45 % d'erreurs d'accord portent sur des verbes ayant une fonction syntaxique similaire : à savoir les verbes (i) 'to have' (x40) ; (ii) 'to be' (x31) ; et (iii) 'to do' (x16). De plus, notons que pour le cas (ii) un tiers des occurrences portent sur des constructions de type existentiel et que les cas (i) et (iii) portent exclusivement sur l'usage relevé avec le présent de l'indicatif pour l'auxiliaire 'do' et le 'present perfect' pour les cas de 'have'.

53. This cartoon shows one of the many things that **have* has been made to […]. (txt_016_sm1)
54. There **is* are a lot of reasons for the glass ceiling […]. (txt_018_sm2)
55. In fact, the international system **don't* $does not$ care about […]. (txt_036_sm1)

(iii) L'accord entre sujet et verbe lexical. (*t107 occurrences*) Les 55% des erreurs d'accord observées entre le sujet et le verbe lexical se limitent à 81 tokens ou items individuels. C'est-à-dire que le choix du verbe n'est pas pertinent puisque que chacun ne se répète qu'à une fréquence de 1,3 fois. Ce qu'il faut donc retenir est que l'accord, indépendamment du choix verbal, pose un problème assez récurrent. Deux exemples sont fournis ci-après.

56. It shows that in appearance the measures already taken **seems* $seem$ to be […]. (txt_019_sm1)
57. That is to say, it is a program which **penalise* $penalises$ or **encourage* $encourages$ countries […]. (txt_041_sm1)

Etant donné les écarts considérables observés dans la fréquence des erreurs du groupe verbal – notamment en termes de statistiques brutes -, il ne nous semble pas judicieux de détailler les autres catégories de façon aussi précise. En effet, après les erreurs de choix lexicaux et les erreurs d'accord, en termes de fréquence, il y a des erreurs portant sur (i) la formation d'un groupe verbal avec un modal (x41) ; (ii) l'emploi de la négation, notamment l'absence d'une partie de la construction 'do not' (x29) ; (iii) la formation des différents 'perfect tense' (x28) ; (iv) la construction du passif (x27) ; et (v) la construction des phrases interrogatives, bien entendu où l'élément épineux se trouve être le pronom interrogatif et l'ordre inversé des items du groupe verbal (x26). Ce qu'il faut retenir de ce fait n'est pas la faible fréquence des autres catégories, par

rapport aux deux premières, mais plutôt le nombre considérable de catégories différentes où des erreurs sont repérées et le nombre quasi-homogène entre elles. Soulignons par ailleurs qu'il y a d'autres catégories où des erreurs sont recensées selon le schéma d'UAM. Toutefois les fréquences relevées varient de 17 à 1, ces catégories ne sont donc pas rapportées ici.

En bref, malgré la multiplicité des catégories individuelles « à problème », avec sans oublier, bien entendu, les effectifs en dessous de 50, nous pensons, en tant qu'enseignant, que ces catégories sont celles qui présentent un plus grand risque de fossilisation chez les apprenants. En effet, si l'on considère à titre comparatif que le noyau ou le n-tête du groupe nominal comprend 393 erreurs comptabilisées, l'ensemble des erreurs du groupe verbal porte en quelque sorte sur son noyau et à ce titre le nombre est deux fois supérieur à ce qui est observé sur le n-tête. Vu de cette manière, cela permet de relativiser le nombre de sous-catégories « d'erreurs verbales » fournies par le schéma d'annotation. Il faut donc comprendre que, sans être alarmiste, le nombre d'erreurs différentes dans le groupe verbal est sujet à réflexion – tant pour le linguiste que pour le didacticien sur le terrain.

5.1.2.3 « Prep-phrase-error »

Passons maintenant aux erreurs de préposition qui, selon les annotations obtenues, peuvent être réparties en quatre catégories différentes. Ces catégories sont les suivantes : (i) le choix lexical erroné ; (ii) l'inadéquation du type de complément ; (iii) l'usage non requis ; et (iv) l'absence d'une préposition requise. Il convient également de préciser que parmi les cas annotés dans la catégorie grammaticale, les erreurs de préposition sont troisième en termes de fréquence après celles relevées dans le syntagme nominal et celles dans le syntagme verbal.

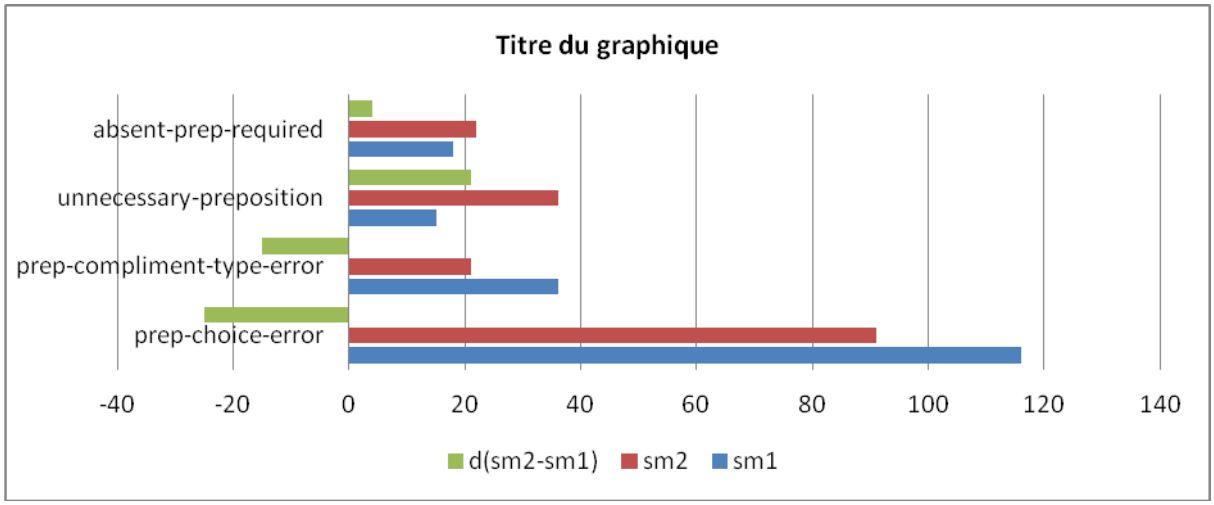


Figure 28 : La comparaison entre semestres des erreurs prépositionnelles

(i) le choix lexical erroné. (t*207 occurrences*) Il s'agit principalement ici de l'usage d'une préposition qui ne convient pas à la situation d'énonciation : à savoir (i) par exemple dans la majorité des cas soit le choix prépositionnel est régi et donc « figé » par rapport au verbe qu'il précède – ce qui est le cas du verbe intransitif dans l'exemple 58 ; soit (ii) le choix n'est pas fixé par le verbe mais est plus ou moins régi en fonction du rapport sémantique entretenu avec les différents éléments syntaxiques que l'on cherche à rapprocher.

58. The jobs that depend **of* on exportations in each country are more and more numerous […]. (txt_044_sm1)
59. We need higher standards **about* $concerning$ green house gases emissions because […]. (txt_003_sm1)
60. Indeed, Abdullah Gül's speech sounds like a way to defend the British *government[sic]* action **of* $over$ the last decades. (txt_049_sm1)

(ii) l'inadéquation du type de complément. (t*57 occurrences*) S'ensuit également l'erreur qui survient quand la préposition exige un certain type de complément : à savoir un syntagme nominal d'un type précis ou un gérondif comme dans les cas ci-après.

61. Nowadays, we can't go anywhere, watch anything on TV or listen to the radio without **see* $seeing$ or **heard* $hearing$ an ad or a commercial. (txt_046_sm1)
62. Certain companies are aware **to pollute* $of polluting$, so they try to do something for the environment [...]. (txt_003_sm1)

(iii) l'usage non requis. (t*51 occurrences*) Comme le titre l'indique, la préposition n'est pas requise et ne joue par conséquent aucun rôle dans la phrase. Ce cas survient majoritairement quand un verbe est employé comme un intransitif alors que celui-ci est en fait bien transitif.

63. […] indeed, we need to guarantee **to* $~~to~~$ the future generations that they will have enough resources to live […]. (txt_031_sm1)
64. If we remove advertising, people would buy what they really want to buy and maybe it would benefit **for* $~~for~~$ the economy. (txt_029_sm1)

(iv) l'absence d'une préposition requise. (t*40 occurrences*) Il est essentiellement question ici de verbes intransitifs qui ne sont pas suivis d'une préposition, d'où le signalement d'erreur.

65. The food industry uses advertising a lot […]. It aims at appealing **Ø* to people so that they consume your product […]. (txt_057_sm1)
66. For example, companies can search **Ø* for new sources of energy and resources to *remplace[sic]* oil. (txt_009_sm1)

Il nous semble a priori que l'ensemble des erreurs de prépositions identifiées relève d'un problème lié au phénomène de transfert négatif du français à l'anglais et traduit de ce fait un besoin réel qui doit dépasser l'apprentissage des prépositions comme étant du lexique à apprendre, comme tous les autres mots – notamment si l'on veut remédier convenablement au problème.

5.1.2.4 « Clause error »

(Clause-error → erreurs de proposition simple 'clause simplex')

Il s'agit d'indiquer sous l'étiquette 'clause error' les éléments qui obstruent non-seulement un niveau syntaxique en tant qu'items lexicaux isolés, mais tout particulièrement ceux ayant une fonction proprement propositionnelle : à savoir les éléments permettant à la 'proposition' d'exister en tant qu'unité syntaxique complexe ou simple. Autrement dit, l'objectif ici est de signaler les points les plus problématiques au niveau de ce que l'on appelle communément les propositions principale, indépendante et subordonnée. Les erreurs en question sont mises en avant dans le tableau ci-après et concernent principalement les ajouts, les connecteurs, les pronoms relatifs et ainsi de suite.

les sous-catégories 'clause-error'	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
adjunct-error	88	76	**164**	-12	-13,6
connector-error[93]	56	48	**104**	-8	-14,2
transitivity-error	23	22	**45**	-1	-4,34
relative-clause-error	14	18	**32**	+4	+28,5

Tableau 26 : Répartition des erreurs dites de « clauses-errors »

(i) L'erreur d'ajout. (t*164 occurrences*). Pour rappel, l'ajout *'adjunct'* au sens large est un complément d'information ou tout autre élément occupant une fonction syntaxique accessoire et donc non-essentielle dans une proposition. Cela dit, les ajouts occupent de manière générale une fonction adverbiale. Il est ressorti de nos annotations que les erreurs d'ajout peuvent être classifiées de trois façons distinctes : (i) il s'agit, tout d'abord, de l'agencement de l'ajout par rapport au verbe et à l'objet ou au complément dans la phrase ; (ii) il est ensuite question du choix lexical inapproprié : à savoir soit (a) l'élément en position d'ajout n'appartient pas à une classe grammaticale permettant d'occuper cette fonction, par exemple en étant un syntagme nominal au lieu d'un syntagme prépositionnel comme ci-dessous (b) soit le choix lexical n'est tout simplement pas sémantiquement approprié, et ce, au vu du contexte informationnel.

[93] Ces erreurs sont présentées en section 6.1.3

67. […] he reported that MBA classes are not *_enough_ connected $connected enough$ with the world of work. (txt_007_sm2)
68. However, […] in the society a gender gap which has significantly _rose[sic]_ *_Ø_ $over$ _this[sic]_ past years. (txt_060_sm1)
69. Many students choose subjects as drama and theatre *_in spite of_ $instead of$ French or German. (txt_036_sm2)

(ii) L'erreur de transitivité. (t*45 occurrences*) Ce qui est signalé ici concerne aussi bien les éléments de transitivité en termes de linguistique systémique qu'en grammaire traditionnelle. En effet, il est question d'erreurs touchant les objets directs ou indirects, les compléments d'objet ou encore des cas où le rôle de sujet qui n'est pas correctement assuré. Parmi les différents cas relevés, il est notamment question (i) de complément obligatoire absent (x18) ; (ii) de sujet obligatoire absent (x9) ; et (iii) d'erreur de construction ditransitive (x7).

70. Policies like carbon offsetting enable *_Ø_ $countries$ to compensate or in the best case offset emissions which […]. (txt_052_sm1)
71. In addition, if all of a sudden the world became advertising free *_Ø_ it will directly _rise[sic]_ advertising and also [...]. (txt_026_sm1)
72. They *_propose their students to do training sessions_ $propose training sessions to their students$, they organize conference [...]. (txt_014_sm2)

(iii) L'erreur dans la construction d'une proposition relative. (t*32 occurrences*) Il est essentiellement question ici du pronom relatif qui (i) est absent ; (ii) ne permet pas de reprendre le groupe nominal auquel il est censé se substituer ; voire (iii) n'appartient pas à la 'bonne catégorie' des pronoms.

73. […] but, in the world of work, it is always _a work[sic]_ with colleagues, *_Ø_ $where$ communication is important. (txt_038_sm2)
74. The companies want to find people *_which_ who will be able to help them […]. (txt_008_sm2)
75. To conclude, I would say that fighting for women's right has begun with the access to the same education, *_what_ $which$ represented a huge shift in mentalities, and nowadays the fight is not over. (txt_042_sm2)

(Clause-error → erreurs de clause complexe)

Il est important de bien distinguer ici entre la « clause simplex » et la « clause complex ». Pour rappel, la première renvoie de manière générale à une proposition indépendante tandis que la

deuxième renvoie à des phrases avec au moins une proposition dépendante et une proposition principale. En effet, comme le tableau 27 ci-dessous l'indique, cette sous-catégorisation avec quatre groupes représente à elle seule un total de 307 erreurs. Les deux premières catégories ont des titres explicites, donc un seul exemple sera fourni par catégorie. La troisième catégorie n'est pas statistiquement significative contrairement à la quatrième qui s'est montrée particulièrement riche en éléments décisifs. Cette dernière sera, de ce fait, regroupée et explicitée davantage avec des erreurs de cohérence, dans une sous-section ultérieure.

les 'clause-complex-error'	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
missing-clause-conjunction	46	31	**77**	-15	-32,6
incorrect-clause-conjunction	23	21	**44**	-2	-8,7
incorrect-clause-type-subordinate	1	0	**1**	-1	-100
incorrect-tense-in-clause-complex	93	92	**185**	-1	-1,07

Tableau 27 ; Répartition des erreurs de « clause-complex »

76. […] thus, compared to boys, girls are more sensitive, *Ø and are more interested in humanitarian activities. (txt_028_sm2)
77. Greece, which is the *more[sic]* indebted country in the Union is today followed by Spain, Ireland, Portugal **or* and Italy. (txt055_sm2)

5.1.2.5 « Les autres erreurs grammaticales »

Sont regroupées dans le tableau 28 ci-dessous les erreurs qui ont été annotées, mais dont la fréquence a été relativement infime par rapport aux autres catégories. Elles ne sont, par conséquent, citées qu'à titre d'illustration.

les autres erreurs grammaticales	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
adjectival-phrase-error	66	50	**116**	-16	-24,2
adverb-error	33	22	**55**	-11	-33,3
special-structure-error	12	24	**36**	12	200
other-grammatical-error	5	4	**9**	-1	-20

Tableau 28: Répartition des autres erreurs grammaticales

Il s'agit donc :

(i) d'erreurs de phrase adjectivale dont le schéma d'annotations d'erreur d'UAM avait prévu huit sous catégories. Le tableau 28 signale donc par exemple que 66 erreurs en sm1 ont été distribuées sur 8 catégories, avec celles portant sur le choix lexical inapproprié enregistrant 29 des 66 items étiquetés. Cela étant, le faible total relevé dans les sept autres catégories suggère une certaine irrégularité par rapport au corpus global.

(ii) d'erreurs d'adverbe dont le caractère peu fréquent est manifeste, avec seulement 33 occurrences comptabilisées au premier semestre. Encore une fois il s'agit majoritairement de choix lexical erroné ou d'un adverbe redondant, sans fonction utile dans la phrase.

(iii) d'erreurs portant sur une construction particulière, comme dans la formation du comparatif en anglais : avec la structure '*more + adjectif + than'*, ou *'more than + np'*. La plupart des erreurs relevées dans ce contexte portent sur *than* dans la mesure où un autre item lexical (comme '*as*' ou '*that*') se trouve être substitué à '*than*'. Toutefois, ces cas sont très inhabituels et ne mériteraient qu'un bref signalement en classe de langue, si l'enseignant venait à en repérer lors des différentes productions en anglais langue étrangère.

(iv) d'erreurs diverses qui ne rentrent pas dans les catégories grammaticales prévues par le schéma d'annotation. Il est question ici d'exemples d'une non-maîtrise flagrante à plusieurs niveaux, ce qui fait que l'élément en question peut être signalé dans de nombreuses catégories sans que l'on puisse identifier un seul point grammatical comme en étant la cause principale.

Ce que l'on doit retenir dans ces deux dernières catégories d'erreurs grammaticales (cf. les erreurs de proposition et les erreurs diverses) c'est que les erreurs ne sont pas forcément là où on les attend. C'est notamment le cas des erreurs de type « ajouts » qui comptabilisent un total de 164 occurrences avec une distribution quasi-homogène dans l'ensemble de la phrase : à savoir aussi bien à la périphérie gauche et droite ou en position médiane. A cela s'ajoute diverses erreurs grammaticales qui se sont illustrées par leur caractère tout à fait singulier, ce qui fait que leur description semble traduire des erreurs humaines à la fois individuelles et ponctuelles plutôt que des erreurs d'apprentissage ou une non-maîtrise linguistique d'un ensemble de participants plus large. Nous pensons donc que ces erreurs ne nécessiteraient pas un temps de correction considérable, mais pourraient faire l'objet d'un rappel simple si l'erreur produite est identifiée en classe de langue.

5.1.3 Les erreurs de ponctuation

Hormis les problèmes d'usage stylistiques, les items qui nous intéressent ici sont uniquement ceux qui engendrent des questions d'agrammaticalité et d'usage conventionnel en langue anglaise. Dans certains cas, l'emploi est tout à fait acceptable dans la langue source de la plupart de nos sujet-participants – à savoir le français, mais ne l'est pas dans la langue cible. En effet, si l'on se réfère

au tableau 29 ci-dessous chaque apprenant ferait plus de trois erreurs de ce type, sans compter bien entendu celles qui relèvent d'erreurs d'inattention ou de style. Toutefois, soulignons une diminution globale de 37%, ce qui est un « bon signe » pour l'acquisition de l'usage de ces marqueurs que l'on pourrait qualifier à la fois de typographiques et orthographiques.

les erreurs de ponctuation	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
unnecessary-capitalisation	35	18	53	-17	-48,6
capitalisation-required	25	19	44	-6	-24
punctuation-inserted-not-required	26	7	33	-19	-73,1
punctuation-required-not-present	20	14	34	-6	-30
wrong-punctuation	11	15	26	4	36,36
missing-space-separator	7	5	12	-2	-28,6

Tableau 29 : Répartition des erreurs de ponctuation

(iv) Majuscule présente non-requise. Pour ce qui est de la première sous-catégorie, on a affaire à des « majuscules » là où on n'en n'a pas besoin. Cela peut être anodin pour certains mais ne l'est pas dans le genre et le registre textuels auxquels les écrits doivent être intégrés : à savoir l'écrit universitaire avec ses exigences propres. Pour parler de la nationalité en français, par exemple, il est possible d'utiliser aussi bien une majuscule qu'une minuscule : le premier est valable uniquement quand celui-ci occupe la fonction de n-tête (les Chinois) et le deuxième quand celui fonctionne en tant que modifieur (les étudiants chinois). En anglais, par contre, les deux cas de figure requièrent une majuscule. Les exemples qui suivent témoignent d'un usage de majuscule quand celle-ci n'est pas requise.

78. The world has to tackle the **Challenge* $challenge$ posed by **Climate Change* $climate change$. (txt_056_sm1)
79. This is the reason why **Degrees* $degrees$ need to be more connected with the real world [...]. (txt_061_sm2)

(v) Majuscule absente nécessaire. La deuxième sous-catégorie va de pair avec la première : il est question ici au contraire d'usage requis d'une majuscule.

80. That is why many countries, during the **kyoto's* $Kyoto's$ protocol [...] (txt_041_sm1)
81. Then **european* $European$ politicians and economists understood quickly that a new single economy [...] (txt_012_sm2)

(vi) Problème de ponctuation. (t*106 occurrences*). Les troisième, quatrième et cinquième sous-catégories (*t93* occurrences) du tableau 6.1.5-1 renvoient à des marques strictement typographiques. Trois exemples sont fournis ci-dessous : (i) pour le premier, il s'agit de

l'emploi d'une ponctuation non-requise ; (ii) de l'absence d'une ponctuation requise ; et enfin (iii) tout simplement du choix erroné de ponctuation pour le troisième.

82. Indeed, a debt crisis started*[,]* that Europe was not ready to face. (txt_013_sm2)
83. [...] it is a real pool of sovereignty, pooling their debt together*[Ø]* $together, their$ their economic and above all fiscal policy so that the market cannot find weaklings in Europe. (txt_043_sm2)
84. Auto-exclusion is also a reason why they do not get into the top jobs. Women have to present themselves as strong personalities*?* They need to form unions [...]. (txt_052_sm2)

(vii) Problèmes d'espacement. Cette catégorie recense les items, peut-être les moins graves, à cheval entre les erreurs orthographiques et les erreurs typographiques. En effet, on identifie (i) dans un premier temps les cas d'emploi par exemple de deux mots distincts en anglais écrits comme une seule unité lexicale (t*12* occurrences) et (ii) et par extension, l'inverse du premier cas de figure - à savoir l'écriture d'un seul item lexical en tant que deux items distincts (t*19* occurrences). Notons que ces derniers n'ont pas été comptabilisés dans le tableau ci-dessus et se limitent à deux items précis : 'another', 'moreover' et 'throughout'.

85. [...] electric cars or hybrid ones are going, in the future, to replace **gasguzzlers* $gas guzzlers$ which are now the main polluters in the United States. (txt_042_sm1)
86. All the political leaders know the solution but it is hard for them to accept **an other* $another$ sacrifice. (txt_037_sm2).

Si l'on prend le total des occurrences observées dans le tableau 29 et les 19 occurrences supplémentaires ajoutées au point quatre (iv) « problèmes d'espacement », on obtient un total de 221 erreurs de ponctuation. Ce chiffre laisse entendre que les questions de ponctuation, souvent écartées dans l'enseignement d'une langue étrangère, ne sont pas si anodines et mériteraient même que l'on y accorde une attention particulière.

5.2 Le schéma expérientiel : problème de transitivité

L'objectif principal de cette section est de fournir un bref aperçu des erreurs identifiées dans notre deuxième couche ou niveau d'annotation. En effet, dans un premier temps les items ont été annotés et présentés en section 5.1 par rapport aux différentes fonctions lexico-grammaticales exercées et sur lesquelles des erreurs ont été observées. Il est maintenant question d'examiner ces mêmes items qui ont été ré-étiquetés par rapport aux rôles sémantiques qu'ils jouent au niveau de ce que l'on

nomme la métafonction expérientielle et tout particulièrement la transitivité[94]. Rappelons à titre accessoire que ces rôles sémantiques concernent uniquement les trois rôles dits de (i) procès, (ii) de participants et (iii) de circonstances et que tous les trois demeurent un phénomène intrinsèquement lié à la structure propositionnelle ou 'clause' en anglais.

Les résultats présentés dans cette section permettent, par conséquent, d'avoir une idée de la répartition globale des éléments annotés à ce niveau, sans pour autant entrer dans une analyse approfondie[95] de l'implication desdits résultats. L'objectif sous-jacent est d'abord de regarder quels composants, au niveau propositionnel (clause), posent le plus de problèmes et ensuite d'examiner les constructions sémantiques créées par chacun de ce composant. Comme nous l'avons expliqué dans le chapitre IV, le raisonnement derrière cette double annotation est notamment d'étudier la même occurrence erronée sous de multiples angles. De plus, le nombre de ré-étiquetages obtenus à ce niveau d'annotation n'est pas identique à ceux obtenus précédemment. Ceci est dû à la fois au fait que certaines erreurs annotées en section 5.1 n'entrent pas dans le cadre d'un des trois composants ou rôles mentionnés ci-dessus ou s'étendent à plus d'un composant à la fois. A titre d'illustration ni les erreurs de ponctuation ni les erreurs purement lexicales ne peuvent être ré-annotées dans cette section. Le tableau suivant montre donc les résultats obtenus.

les erreurs de transitivité	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
process	472	434	906	-38	-8,05
participant	496	530	1026	34	6,85
circumstance	326	388	714	62	19,01
total	1294	1352	2646	58	4,48

Tableau 30 : Répartition des erreurs de transitivité

5.2.1 Les erreurs de procès

Comme nous avons pu observer dans le tableau 30 précédent, les erreurs portant sur le procès comptabilisent un total de 906 occurrences dans nos deux semestres d'étude, avec une diminution relative de 8 % en sm2. Toutefois un regard critique sur l'ensemble des catégories regroupées à ce niveau permet de mieux situer les types de procès précis qui posent problème à nos apprenants francophones. Ce regard critique est apporté dans la figure 29 où l'on peut constater qu'un type de procès sur les six établis regroupe à lui seul 48% (437/906) de l'ensemble des erreurs relevées sur les procès ou groupes verbaux : à savoir le procès matériel que l'on pourrait désigner intuitivement

[94] Cf. section 3.2.5 pour un rappel sur la transitivité et plus précisément la métafonction expérientielle.
[95] Cf. les sections 5.1 et 5.2.4 pour une l'analyse détaillée de ces résultats, avec notamment un regard croisé sur l'ensemble des résultats obtenus à partir des différents niveaux d'annotation.

comme étant le plus fréquent dans les écrits du type universitaire. Si l'on y ajoute les erreurs observées sur le procès dit relationnel, le pourcentage est désormais supérieur à 70% (643/906).

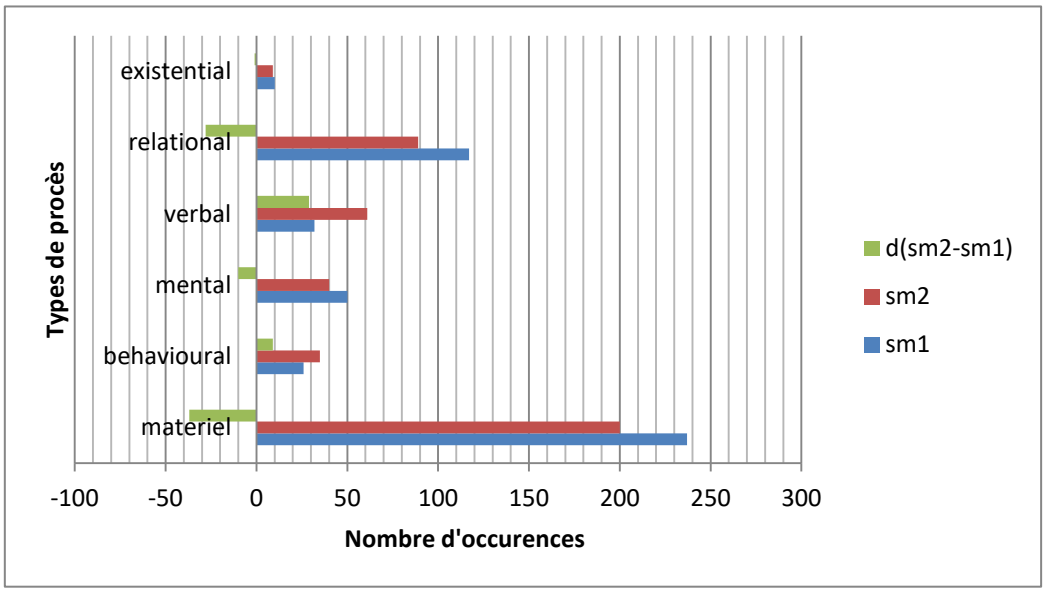


Figure 29 : L'écart relatif des erreurs de procès entre les semestres

En outre, la variable temporelle ou l'écart chiffré observé (d, en vert) entre les deux semestres montre que la diminution n'a été ni proportionnellement constante ni holistique par rapport aux chiffres initiaux obtenus en sm1. En effet, 2 des 6 procès ont enregistré une augmentation d'erreurs en sm2, avec environ 90% de plus pour les erreurs portant sur les procès verbaux et 35% de plus pour les erreurs identifiées dans les procès comportementaux ou 'behavioural'. La valeur réelle de l'étiquetage « erreurs de procès » ne devient de ce fait facilement compréhensible, voire explicite que lorsqu'elle est contrastée avec (i) les 14 sous-catégories précédemment identifiées dans le syntagme verbal du premier schéma d'annotation et (ii) les autres catégories, comme certaines erreurs propositionnelles (clause) ou de cohérence (cf. section 6.1.1.2), où les temps et les aspects verbaux sont remis en cause.

Ces deux graphiques indiquent tout simplement par conséquent que l'ensemble des phénomènes linguistiques comptabilisés à l'intérieur de chaque groupe de procès sont bien variés et distincts les uns des autres. Toutefois, indépendamment de l'angle contrastif qui sera davantage développé dans le chapitre VII, la figure 29 ci-dessus met indiscutablement en avant le fait que certains types de procès posent plus de problèmes que d'autres. Ce qui en fait un constat non négligeable qui, à notre connaissance, n'a pas encore été identifié en tant que tel jusqu'ici.

5.2.2 Les erreurs de participants

Le signalement dans cette section renvoie à la distribution ou au comportement, pour ainsi dire, des items erronés selon leurs rôles de participants dans une phrase donnée. Rappelons que ne peuvent être participants que les syntagmes nominaux qui réagissent par rapport à un noyau verbal ou un 'procès'. La figure ci-dessous donne un aperçu de la répartition des erreurs identifiées sur des éléments jouant le rôle de "participants".

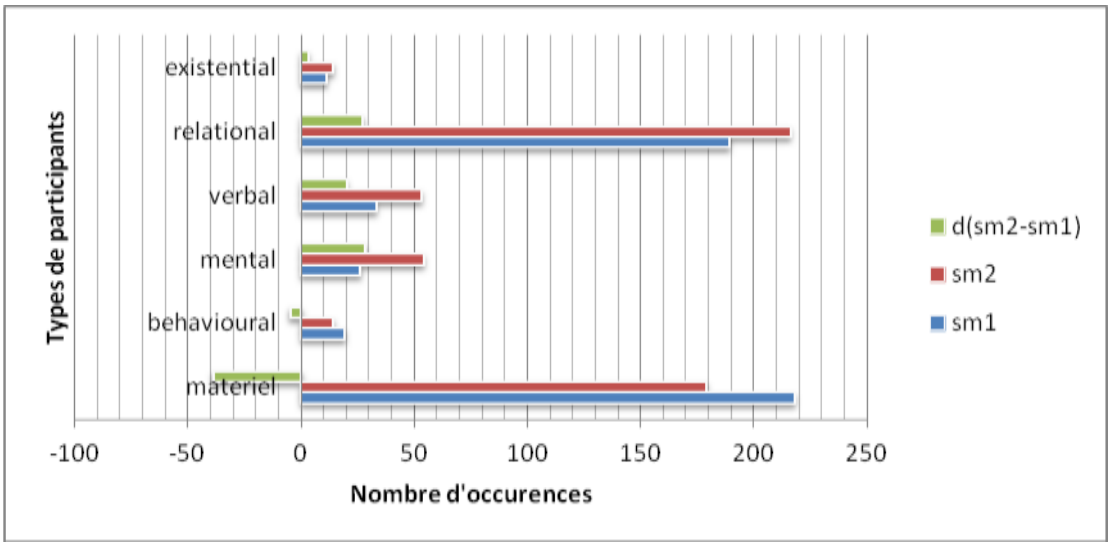


Figure 30 : L'écart relatif des erreurs des participants entre les semestres

Il en ressort, comme ce fut le cas des erreurs sur le procès, que deux catégories sur six regroupent plus de trois quarts (78%) des items identifiés à ce niveau: à savoir les participants relationnels (405/1026) et matériels (397/1026). Autrement dit, les participants renvoyant à des liens relationnels en tant qu'éléments d'identification, d'attribution ou de possession pour le premier cas de figure; et des renvois à des éléments concrets pour le second. Un point de divergence notable entre ces deux catégories reste le fait que d'une part, les erreurs identifiées dans les regroupements de participants matériels ont diminué au deuxième semestre de 18% tandis qu'elles ont bondi de 14% chez les participants relationnels, d'autre part. L'ensemble de ces constats n'est pas surprenant en soi – et ce, étant donné le fait que les procès matériels et relationnels nous semblent être les plus fréquents en anglais général et aussi dans les rédactions de type universitaire et par conséquent, les participants correspondants sont en principe aussi fréquents que les procès auxquels ils sont attachés. Les erreurs relevées ne peuvent donc être que proportionnelles aux différents moyens linguistiques employés.

5.2.3 Les erreurs de circonstance

Tout d'abord, rappelons que ce qui est considéré et annoté comme une erreur de circonstance englobe toute erreur identifiée dans les syntagmes prépositionnels et adverbiaux. Dans de nombreux cas, ces erreurs peuvent être considérées comme non indispensables structurellement, tant en termes de syntaxe que de sémantique et peuvent donc être étiquetées en tant qu'erreur d'ajout (adjunct errors). Cependant, par souci de brièveté, nous utiliserons uniquement la première étiquette telle que mentionnée précédemment dans le chapitre III.

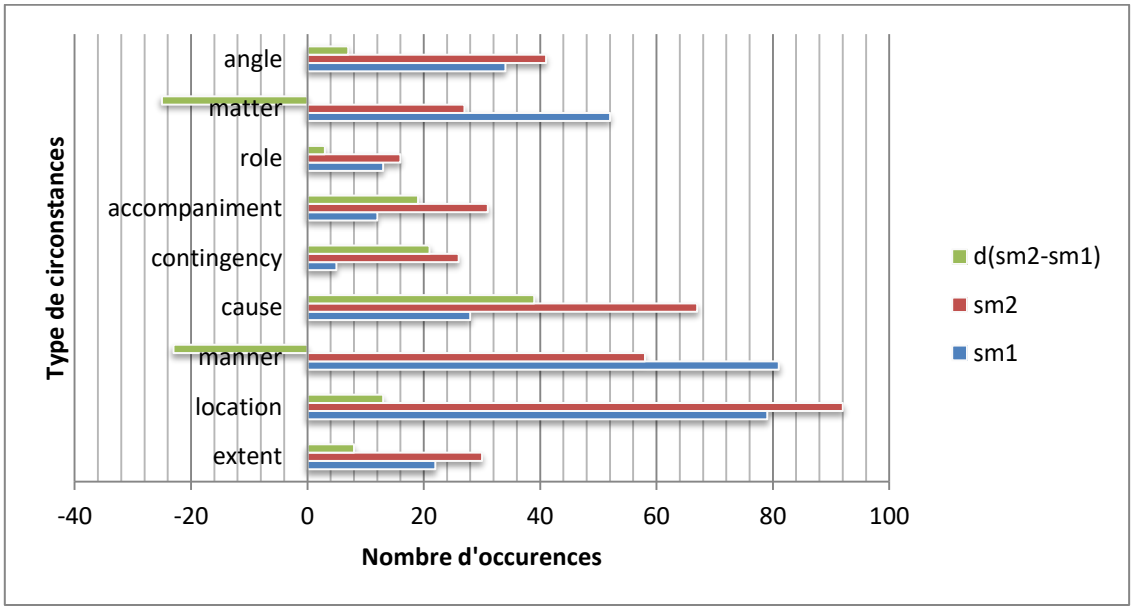


Figure 31 : L'écart relatif des erreurs circonstancielles entre les semestres

Dans la figure 31 ci-dessus, la première chose que l'on peut observer est la distribution fluctuante. Toutefois un examen plus attentif montre que deux catégories sur neuf représentent 43% des annotations globales, tandis que les sept autres catégories d'erreurs de circonstance sont plus également réparties entre elles. Les deux catégories qui semblent être les plus problématiques pour nos apprenants sont les circonstances de 'location' et de 'manière'. La première concerne les ajouts spatio-temporels et se réfère plus précisément à des questions de « quand » et « où », tandis que la seconde catégorie se réfère généralement à des descriptions de degré, de qualité et de comparaison, etc. Des exemples concernant deux cas sont fournis ci-dessous.

(viii) Circonstance de 'location'. (t*171 occurrences*) Il s'agit ici d'une remise en cause de la l'acceptabilité et de la précision spatio-temporelle.

87. Even if much progress was made **during the last years* $over the last years$, we have *still[sic]* much progress to achieve. (txt_019_sm2)

88. There is a contradiction in this issue because women's education rate *show[sic]* that they study, they even go **at* to *the university*. (txt_022_sm2)

(ix) Circonstance de manière. (t*139 occurrences*). Il est question de répondre à la question 'comment' ou d'apporter des précisions en termes de degré et de qualité, etc.

89. [...] there was a crisis in the United States [...]. **Very fastly* $Very quickly$, the world was impacted. (txt_014_sm1)
90. [...] and knowing that this system of carbon offsetting is not **enough efficient* $efficient enough$ [...]. (txt_035_sm1)

Ce qui est également notable dans la figure 31 est que, contrairement à la tendance générale de diminution – et ce, aussi petite soit-elle dans la plupart des erreurs de procès et de participants – les erreurs de circonstances enregistrent une augmentation globale de 20%. De plus, uniquement deux catégories enregistrent une baisse au deuxième semestre par rapport au premier, tandis que les sept autres catégories individuelles enregistrent une augmentation allant de 16 à 420%. Ce surcroit d'erreurs signifie que le problème gagnerait à être traité de manière ciblée dans toute classe de langue étrangère – indépendamment du degré de spécialité.

5.2.4 Le bilan des annotations du schéma expérientiel

Ce niveau de ré-annotation a fourni un moyen de visualiser les erreurs commises par nos apprenants, en les mettant en relation directe avec la structure expérientielle de base de la phrase anglaise. En effet, ces ré-annotations permettent à la fois (i) une analyse globale (dite « coarse-grained ») des éléments erronés en fonction de leurs rôles de participants (PR1 et PR2), de procès et de circonstance ainsi que (ii) une analyse plus fine où les quatre catégories précédemment identifiées peuvent être davantage approfondies. Le résultat global du schéma expérientiel est illustré ci-dessous de manière chiffrée dans le tableau 31 dans lequel l'analyse globale et l'analyse approfondie sont croisées. On peut alors observer les éléments au niveau expérientiel qui s'avèrent les plus épineux pour nos sujet-participants.

les erreurs	**PR1**	**Pr.c**	**PR2**	**C**
materiel	109	437	288	714
behavioural	11	61	22	
mental	21	90	59	
verbal	15	93	71	
relational	166	206	239	

existential	0	19	25	
total	322	906	704	714

Tableau 31 : Répartition chiffrée des erreurs selon le schéma expérientiel

En effet, le tableau ci-dessus rassemble en effet en quelque sorte les grandes lignes des trois sous-sections précédentes : à savoir de *5.2.1* à *5.2.3.* Pour faciliter la compréhension du tableau, on pourrait dire que *PR1* coïncide avec le sujet grammatical, *Pr.c* avec le procès ou groupe verbal, *PR2* avec l'objet ou le complément du verbe en grammaire traditionnelle et *C* correspond enfin aux ajouts non indispensables à la fois sur le plan syntaxique et sémantique. Cela dit, ce tableau fournit, de ce fait, un aperçu dans lequel on identifie non seulement la valeur syntaxique mais également la valeur expérientielle qui pose problème. Ainsi, pourrait-on dire que l'analyse expérientielle est à mi-chemin entre une analyse syntaxique et une analyse plus large dite du discours puisqu'on entre dans le sens que l'apprenant cherche à créer en identifiant les moyens - non seulement syntaxiques mais également sémantiques - qui lui ont fait défaut.

Il est également possible d'illustrer ces mêmes résultats sous un autre format visuel. Les données chiffrées sont certes importantes mais le graphique ci-après facilite encore davantage la compréhension des items annotés à ce niveau. Il montre en effet sous forme d'histogramme les résultats du tableau 31, permettant ainsi de mieux comprendre la répartition des trois éléments clés de la phrase principalement par rapport à la typologie du procès.

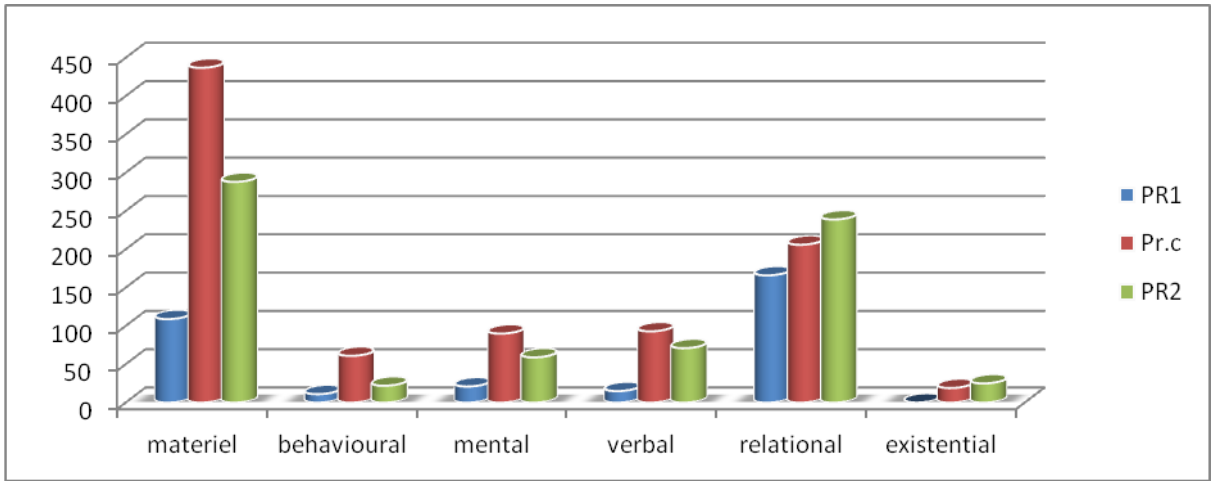


Figure 32 : Mise en rapport graphique des erreurs au niveau expérientiel

Cette nouvelle mise en perspective (i) présente la tendance – que l'on pourrait qualifier de générale – en termes de fréquence réelle des différents emplois en anglais ; (ii) elle souligne le pourcentage que chaque catégorie représente par rapport à l'ensemble (et par conséquent une probabilité que cette tendance soit généralisable avec 12 % d'erreurs possibles et identifiables dans la fonction de

participant 1 (PR1), 34% dans la fonction de procès (Pr.c), 27% dans la fonction de participant 2 (PR2) et à nouveau 27 % dans la fonction de circonstance (C) ; (iii) enfin elle permet également d'indiquer le rapport de force et donc les éléments qui mériteraient le plus de temps de travail en profondeur en classe de langue : et ce, bien entendu dans le but de faire progresser les apprenants.

5.3 Le schéma textuel

Comme nous l'avons signalé antérieurement, les erreurs identifiées par le schéma d'annotation d'UAM ont été ré-annotées deux fois : tout d'abord avec un schéma d'annotation basé sur la métafonction expérientielle et ensuite avec un autre basé sur la métafonction textuelle. Nous allons maintenant détailler succinctement les résultats obtenus lors de la deuxième ré-annotation. En effet, cette deuxième étape fournit un cadre dans lequel les éléments sont - dans une grande mesure - examinés à la lumière de leur structure ou de la composition textuelle plutôt que leur syntaxe générale (cf. section 5.1.2) ou leur métafonction expérientielle (cf. section 5.2). Par conséquent, s'intéresser à la structure textuelle signifie de manière analogue examiner « les blocs de construction » individuels qui sont utilisés pour construire un ensemble textuel plus large. Ce niveau d'annotation vise donc à souligner quels « blocs de construction » sont défectueux et ceux qui minent la construction globale. Les erreurs sont, de ce fait, différenciées en fonction de leur emplacement et des fonctions textuelles : c'est-à-dire, en commençant par le dénominateur commun le plus large, à savoir le thème et le rhème.

textualmf-type	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
Theme	471	407	878	-64	-13,59
Rheme	1444	1303	2747	-141	-9,765
total	1915	1710	3625	-205	-10,7

Tableau 32 : Répartition bipartite chiffrée des erreurs annotées avec le schéma textuel

Le tableau 32 met en avant le rapport de force, pour ainsi dire, qui existe entre la répartition bipartite en « erreurs de thème » et « erreurs de rhème ». Ce qui interpelle le plus dans ce tableau (sans pour autant constituer une surprise) est le ratio 1:4 qui existe entre ces deux derniers. En effet, ce ratio vaut aussi bien pour le premier semestre (sm1) que pour la somme totale obtenue en *n* dans les deux semestres, ce qui signifie par ailleurs que cette répartition constante ne peut pas être un phénomène aléatoire. Ce phénomène peut alors être expliqué en partie par rapport au fait que la structure canonique de la phrase (à savoir SVO ou PR1, Pr.c et PR2) présuppose qu'il y a moins d'éléments en position de (ou à l'intérieur du) thème par rapport au rhème : d'où le ratio 1 : 4. Cependant, précisons d'emblée que nous allons limiter notre analyse détaillée aux erreurs relevées

en position de thème pour deux raisons pratiques : (i) le cadre de la linguistique systémique n'a pas de sous-catégorisation établie pour des unités qui composeraient le "rhème", comme c'est le cas pour le "thème" ; et (ii) nous pensons qu'une analyse croisée suffira pour apporter des éclaircissements aux phénomènes déjà explorés dans les deux schémas d'annotation précédents, notamment lors de la présentation des nombreuses erreurs identifiées en position de rhème.

Les erreurs de thèmes concernent, rappelons-le, tout élément situé à la périphérie gauche de la phrase – et plus singulièrement ce qui est placé avant le procès ou ce que l'on désigne traditionnellement comme le groupe ou noyau verbal. Les informations fournies dans la figure 32 ne nous renseignent pas nécessairement sur une typologie ou des cas précis, mais indiquent plutôt la position et la fonction textuelle de l'item mis en cause. En bref, si nous devions regarder par exemple les erreurs de 'nominal group simplex' identifiées sous la rubrique « topical », ce que l'on doit comprendre est que 259 sur 878 occurrences erronées (soit 29%) ont eu lieu en position initiale de la phrase dans des groupes nominaux simples : c'est-à-dire, principalement dans un syntagme nominal composé en principe de deux (DET + NP tête) ou trois éléments (DET + Pré-modifieur + NP tête). Ce résultat est d'autant plus frappant quand on compare les syntagmes nominaux complexes qui sont composés de deux syntagmes nominaux simples et qui n'enregistrent quant à eux que 31 sur 878 occurrences erronées, soit 3,5%.

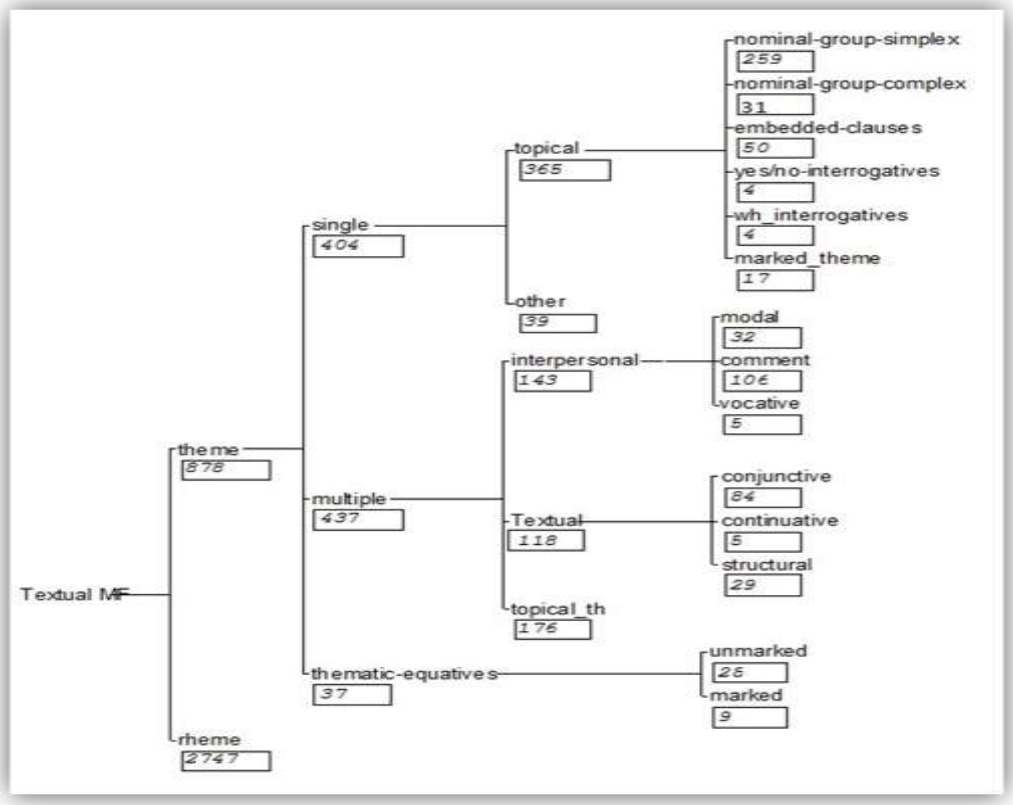


Figure 33 : La distribution schématique des erreurs de thème

Toutefois les erreurs relevées en position de « thèmes – multiples » et notamment celles identifiées dans les rubriques étiquetées 'impersonnel' et 'textuel' s'avèrent très probantes. L'intérêt de ces résultats ne s'explique ni seulement en termes de position et de fonction textuelle (comme ce fut le cas, avec les erreurs 'topicales'), ni simplement en termes de statistiques chiffrées, mais tout particulièrement au vu de la récurrence des cas spécifiques à l'intérieur de ces sous-groupes. De plus, hormis les problèmes d'usage sémantique ou de défigement pour les expressions idiomatiques, certaines erreurs relèvent d'une appréciation plus large du texte. Dans certains cas, des empans textuels larges sont nécessaires pour bien juger de la pertinence de l'étiquetage. Des exemples sont fournis ci-après pour les cas les plus marquants et qui ne nécessitent pas trop de pré-texte ou de post-texte pour la bonne identification des erreurs.

(x) Erreur interpersonnelle (modale). (t*32 occurrences*). L'erreur signalée porte sur la volonté de l'apprenant d'introduire une perspective personnelle dans son discours à travers des éléments de nuances, de probabilités, d'intensité et ainsi de suite. L'erreur survient de ce fait lors que l'élément inséré ne remplit pas son rôle ou n'est pas à sa « place » dans la distribution informationnelle, voire n'appartient pas à la bonne classe grammaticale prévue par le système linguistique.

91. On the other hand, advertising creates new needs through innovation and fashions, which **exactly* $accurately//Ø$ represents characteristics of our modern society. (txt_050_sm1)
92. *In that way[sic]*, summits between France and Germany **especially* have been $exceptionally$ numerous. (txt_054_sm2)

(xi) Erreur interpersonnelle (commentaire). (t106 *occurrences*). Cette sous-section est sensiblement liée à la précédente. Mais afin d'apporter une catégorisation précise, les erreurs qui s'apparentent aux types d'ajout dits de commentaire ou 'comment adjunct' en anglais ont été classées dans cette catégorie. Il s'agit donc plutôt d'éléments évaluatifs, d'opinion, etc.

93. And, **in* as *a matter of fact*, ads are [...]. (txt_046_sm1)
94. **To many ways* $All in all$, "the goal of education is not instruction, but action". (txt_030_sm2)

(xii) Erreur textuelle (conjonctive). (t*84 occurrences*). A ce niveau, il est tout d'abord question dans un texte donné de construire l'environnement textuel dans lequel les « thèmes textuels conjonctifs » signalent comment les différents « blocs de constructions textuelles » sont reliés entre eux. Les éléments erronés peuvent complètement fausser le sens d'un texte s'ils ne sont

pas utilisés à bon escient. Il s'agit donc principalement ici d'erreurs identifiées sur des éléments jouant le rôle de ce que l'on appelle généralement les mots de liaisons.

95. [...] they agree with *environmentalist* to reduce their emissions of carbon and to be careful with their production of waste. **For consequently*, environmentalists suggested projects and [...]. (txt_027_sm1)
96. Namely, it will not always be the case because this action can **in contrary* decrease consumers' consumption and [...]. (txt_026_sm1)

(xiii) Erreur textuelle (structurelle). (t*29 occurrences*). Il est question d'erreur portant sur ce que l'on appelle des conjonctions de coordination ou de connecteurs (cf. page 189).

97. Major countries of the Eurozone cannot afford [...] because it will endanger their own economy. **Then* $Therefore$, countries which are not part of the Eurozone such as Great Britain [...]. (txt_043_sm2)
98. [...] thanks to the new industries and technologies, like internet which helped to spread *formation[sic]* and **so* $to a certain extent$ globalisation. (txt_047_sm1)

En définitive, ce niveau d'annotation a permis d'identifier des problèmes que l'on rencontre souvent dans les productions écrites des apprenants en langue étrangère, sans pour autant arriver à signaler la typologie exacte des occurrences erronées. Toutefois, malgré le fait que cette analyse peut, à première vue, sembler parcellaire voire superficielle, nous soutenons qu'elle se montrerait très bénéfique pour celui qui doit faire face à une classe de langue dans laquelle la syntaxe n'est qu'un aspect secondaire – comme dans le cas des cours d'anglais de spécialité ou les cours de LANSAD. En effet, cette analyse textuelle pourrait être particulièrement utile dans ces contextes où l'objectif principal porte de manière générale sur des compétences proprement rédactionnelles. Cependant nous ne nous attarderons pas davantage sur cet aspect ici mais nous reviendrons sur son utilité didactique dans les chapitres suivants.

5.4 Le bilan des annotations d'erreurs du système linguistique

Les trois étapes d'annotation réalisées sur ce que l'on a nommé les erreurs du système linguistique se sont avérées considérablement riches au niveau de l'identification et la classification des items jugés erronés. Tout d'abord, on a pu annoter des éléments sur plusieurs niveaux avec le schéma d'annotation d'UAM. Et bien que le schéma soit particulièrement détaillé, on a pu remarquer des chevauchements de caractéristiques sur plusieurs étiquetages : par exemple les erreurs proprement lexicales (cf. section 5.1.1) et les erreurs de choix lexicaux liées aux différentes classes

grammaticales (cf. section 5.1.2). L'ensemble de ces items demeurent néanmoins des erreurs mais les délimitations des étiquetages ne sont pas toujours très bien définies[96].

En effet, les erreurs lexicales sont bien plus nombreuses que pourraient laisser croire les catégories initiales prévues à cet effet et tout singulièrement les chiffres récapitulatifs signalés dans le tableau 17. Cet écart s'explique par le fait que la grande catégorie qui « chapeaute » toutes les erreurs lexicales ne renvoie qu'à très peu de sous-types possibles, alors que le nombre de phénomènes existants dépasse largement les catégories prévues. Dans certains cas, des étiquetages qui auraient été mieux « raccordés » à la grande catégorie lexicale sont « rattachés » à la classe grammaticale. Et bien que l'on admette que le choix de la systématicité voudrait que toutes les erreurs soient signalées selon la classe grammaticale à laquelle appartient le terme, la présente configuration pose problème. Nous reviendrons sur ce point en détail dans la conclusion.

Mis à part cet écueil, les nombreuses erreurs annotées avec le premier schéma ont permis d'avoir une vue d'ensemble sur les points les plus épineux par rapport à chaque classe grammaticale – et notamment la fonction syntaxique jouée par les items annotés. Les croisements d'analyse obtenus ultérieurement après la deuxième et troisième ré-annotation ont permis d'aborder les mêmes erreurs sous une perspective différente : à savoir l'identification de la valeur individuelle des items étiquetés par rapport à l'ensemble sémantique et textuel que les sujets-participants cherchaient à créer.

En somme, les trois niveaux d'analyse ont fourni un cadre permettant de montrer qu'il existe des moyens différents de signaler et étiqueter des erreurs : (i) en fonction de la syntaxe des mots utilisés; (ii) en fonction du sens créé par les items dans le contexte d'emploi ; et (iii) en fonction du rôle joué par les items dans la construction d'un ensemble textuel plus large. Chacun pourrait alors librement choisir une des méthodes sus-citées selon l'objectif linguistique et didactique de son projet d'analyse – voire choisir les trois, de façon à apporter un regard neuf et holistique sur les différents types d'erreurs du système linguistique qui continuent d'intriguer des spécialistes en linguistique, didactique et bien des disciplines connexes.

[96] Nous examinerons quelques-unes des conséquences de ces chevauchements dans la section 8.4.2.

(Chapitre VI) Résultats des erreurs textuelles

Dans ce chapitre, il sera question d'examiner dans un premier temps les erreurs qui ont, certes, été annotées lors du premier volet d'annotation, mais qui se sont avérées – pour la plupart – manifestement d'ordre textuel. A travers un examen minutieux de ces premières erreurs textuelles, nous montrons les limites et les chevauchements entre ce qui relève du proprement textuel et ce qui relève du système linguistique de la langue. Dans un deuxième temps, nous discuterons des résultats obtenus lors du deuxième volet d'annotation de notre étude. Ce volet, rappelons-le, ne s'intéresse qu'aux items qui ne peuvent être considérés comme étant des erreurs, non pas au regard du système linguistique lui-même, mais uniquement au vu de l'environnement textuel immédiat. Nous appellerons ces phénomènes des erreurs d'acceptabilité textuelle.

Dans un souci de clarté, il convient de préciser toutefois que seules les occurrences relevées dans des énoncés dépourvus de toute agrammaticalité ont été retenues dans le deuxième volet afin d'éviter au mieux la confusion ou le mélange des genres. En effet, cette sélection est due au fait qu'un nombre élevé d'erreurs d'acceptabilité textuelle a été repéré à travers notre corpus d'étude mais se trouvait dans des phrases où l'ensemble n'était pas grammaticalement acceptable, ce qui aurait pu entretenir une sorte de « double peine » si l'on avait choisi d'annoter les items relevés à la fois pour leur grammaticalité et pour leur acceptabilité textuelle. Le chapitre qui suit est donc divisé en trois sections : la première introduit les erreurs textuelles issues du premier volet d'annotation ; la deuxième explore l'ensemble des erreurs d'acceptabilité textuelle et les examine ensuite à la lumière des métafonctions LSF ; et enfin, la dernière section fait état du bilan de l'ensemble des annotations opérées dans ce volet.

6.1 Les erreurs textuelles du schéma d'annotation UAM (volet 1)

Comme nous l'avons mentionné, certaines des occurrences annotées dans le premier volet d'annotation ne sont pas imputables au système linguistique lui-même. Ces erreurs qui sont - pour la plupart - proprement textuelles ou relèvent des chevauchements entre plusieurs catégories sont présentées ci-dessous.

6.1.1 Les erreurs pragmatiques

Les erreurs relevées ici répondent aux lignes ou 'questions' directrices suivantes : est-ce que les éléments sous étude font ce qu'ils sont censés faire ? Est-ce qu'ils créent convenablement le sens ou l'effet recherché ? Est-ce qu'ils permettent d'établir le lien entre deux ou plusieurs éléments de manière convenable, au vu du contexte à la fois informationnel et textuel ? En effet, les erreurs pragmatiques témoignent d'un frein supplémentaire dans la production écrite en langue étrangère : un frein qui est en quelque sorte aussi assujetti à la grammaticalité. Et ce, parce que la grammaticalité et l'acceptabilité pragmatique vont de pair dans l'appréciation d'un texte comme un ensemble textuel acceptable.

les erreurs pragmatiques	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
cohesion-error	99	82	181	-17	-17,2
coherence-error	33	18	51	-15	-45,5
register-error	94	58	152	-36	-38,3
other-pragmatic-error	0	1	1	1	####
total	226	159	385	-67	-29,6

Tableau 33: Répartition des erreurs pragmatiques

Au vu du tableau ci-dessus, on dénombre un total de 385 occurrences erronées avec 226 au premier semestre et 159 au deuxième semestre, ce qui traduit une amélioration ou une baisse de 29,64%.

6.1.1.1 « Pragmatic-error → erreurs de cohésion »

Est appelé « erreur de cohésion » tout ce qui fait défaut à la liaison structurelle entre plusieurs items ; et ce - qu'elle soit un item lexical individuel ou un empan textuel ou discursif plus large. En effet, le lien cohésif est compris dans notre étude comme étant un dispositif principalement structurel. Les deux cas les plus fréquemment identifiés sont illustrés ci-après. Notons toutefois qu'un contexte textuel plus large est souvent nécessaire pour apprécier les erreurs à leur juste valeur.

(i) Problème de saillance entre référents. (t*149 occurrences*). Cela signifie que l'item choisi ne permet pas d'établir un lien de référence claire. Par exemple, il peut être question de l'emploi

d'un pronom anaphorique (ou anaphorisant) sans qu'un antécédent soit suffisamment saillant pour établir un lien de référence distincte. De manière générale, on peut soutenir qu'il s'agit respectivement d'une tentative d'anaphore résomptive et une anaphore coréférentielle dans les deux exemples suivants.

1. Women even choose to take a part-time job, and are very qualified, at the same time. Moreover, social protection *prevent[sic]* them from hardships at work. They sometimes learn their rights and how legislation works at the university. **This one* $The university$ also delivers new tools, useful on the labour market (for instance how to write a CV, a letter, virtual tools power point, ...). Thanks to the tuitions fees, a lot of universities look for intelligent and *briliant[sic]* students, to help them to climb on the social ladder. (txt_030_sm2)
2. If women don't get more job opportunities it is because of sexism which is not *assumed[sic]* by the employers who *unconsciensly[sic]* use it. Sexism is defined as the fact to *discriminate[sic]* somebody because of his gender. Women are the first victims. It is easy to find employers who do not hire women because they are women and *with it they[?] found too much difficulties[sic]*. For example, **they* $employers/women?$ usually think that **we* $the employers$ cannot give **them* too much *responsability[sic]*, as they are too emotional they won't be able to make important decisions. (txt_047_sm2)

Dans le premier exemple, l'emploi de l'adjectif démonstratif « this one » en tant qu'anaphore résomptive pose plusieurs problèmes d'acceptabilité : mais tout particulièrement, dans ce contexte précis, « this one » ne permet d'établir aucun lien avec « at the university ». Dans le deuxième exemple, le fait d'enchainer des pronoms insuffisamment saillants entre eux crée un effet de confusion. En effet, la dernière phrase – qui comporte pas moins de cinq pronoms – demeure grammaticalement correcte, mais textuellement et sémantiquement confuse.

(ii) <u>Problème d'incompatibilité entre référents.</u> (t*32 occurrences*) Ici, l'erreur apparait quand l'antécédent pour lequel on cherche à créer une référence anaphorique n'est pas compatible par rapport au nombre, genre, voire au type de référence pronominale.

3. Besides women are expected to do the majority of *this[sic]* household tasks, preparing the meal, bringing up children ... So it seems to be difficult for **her* $them$ to balance work and family. (txt_056_sm2)
4. [...] Greece has started to inject **his* its former currency *in[sic] *his* its economy because the leaders fear a bankrun. (txt_058_sm2)

Dans les cas d'incompatibilité entre référents, on observe de manière générale trois cas de figure. L'erreur est (i) uniquement d'ordre textuel ; (ii) uniquement d'ordre linguistique ; (iii) aussi bien d'ordre linguistique que textuel. Ces chevauchements sont discutés dans la section 6.1.4.

6.1.1.2 « Pragmatic-error → erreurs de cohérence »

A l'inverse de la notion structurelle qui prédomine pour la cohésion, la cohérence demeure un dispositif sémantique au service de la distribution et de la continuité informationnelle. Ainsi est considéré comme erreur de cohérence tout ce qui provoque une rupture dans la chaîne à la fois informationnelle et thématique. Autrement dit, tout ce qui fait défaut à une construction discursive destinée à être une unité collective ou un ensemble proprement textuel à part entière.

(iii) Problème de temps grammatical employé. (t*206 occurrences*). Soulignons d'emblée que seulement 21 occurrences ont été annotées dans cette catégorie, et ce dans des phrases dites simples. En effet, l'erreur du temps grammatical survient dans une « phrase » qui est avant tout grammaticalement correcte, mais dans laquelle le temps n'est pas en adéquation avec l'environnement textuel et discursif. Les 185 occurrences supplémentaires – avec le même problème de temps grammatical - ont été identifiées dans les phrases dites complexes. Dans ces cas, un problème de concordance des temps est souvent observé dans les propositions dépendantes.

5. […] half of the candidates of each election in each political organisation must be women. It seems to be a main breakthrough, but *if only[sic]* it **has* worked $only if it works$! (txt_031_sm2)
6. At the beginning the European Union was only an economic union, then it was a single market, and now some countries have the same currency. But it **becomes* $is now becoming$ a political union too. (txt_039_sm2)

(iv) Problème de modal. (t*26 occurrences*). L'erreur portant sur le temps grammatical ne se limite pas aux verbes lexicaux ou aux auxiliaires mais également aux auxiliaires modaux. Toutefois, il faut pouvoir dépasser la phrase isolée – dépourvue de contexte textuel – pour apprécier de manière convenable le problème ici. Et ce, étant donné que l'on peut facilement procéder à des commutations des modaux sans que cela change la structure syntaxique. Cependant, de telles commutations modifient complètement le champ sémantique, d'où l'importance du contexte textuel plus large dans l'appréciation de ces erreurs.

7. As a conclusion, because of mutations of the world of work, degrees must be reformed in order to be more connected with the labour market: the *high education[sic]* system must be considered as a "training ground" with more work placements and oral *assesment[sic]* in the field of languages in degrees. Nevertheless, university **doesn't have to* $should not$ forget that it has to lead to the students' self-*fullfulment[sic]* too. (txt_045_sm2)
8. In *a[sic]* conclusion, I think that the *world[sic]* "collapse" is too strong to qualify the evolution of the global economy in case of an advertising free world. It **must* $might$ provoke strong reactions at first, but money **could* $would$ still lead to power. Companies, that are lobbyists, could ask even more *to[sic]* politicians in order to help *them[sic]*: the economic world would still *stay[sic]* unfair at a national or global scale. (txt_051_sm1)
9. To conclude the cartoon represents our world today. A world where producing from fossil fuels is much more developed than producing from renewable energies as windmill or *soles[sic]* power. This shift must be made by companies and by governments and carbon offsetting is one of the ways by which companies **will* can reduce their greenhouse gases emissions. (txt_034_sm1)

L'exemple (7) met en avant le choix erroné d'une périphrase modale. Dans les nombreux exemples similaires que nous avons pu annoter, l'erreur de cohérence ne porte pas sur un seul point précis. Dans l'exemple (8), l'erreur porte à la fois sur le sens et le type de modal employé. On peut également remarquer un certain nombre de chevauchements portant (i) sur le sens du modal employé ; (ii) sur le temps lié au modal employé ; et (iii) conjointement sur le sens et le temps liés au modal, comme dans l'exemple (9).

(v) <u>Problème de connecteur ou de mot de liaison.</u> Ce point rejoint les 104 occurrences qui ont été signalées précédemment dans 'erreur de connecteur', identifiées précédemment au niveau des 'clause-error' (cf. section 5.1.2.4). L'ensemble des erreurs de ce type sera présenté en détail dans la section 6.1.3.

Ces trois types d'erreurs de cohérence que nous venons d'aborder ne sont pas les seules erreurs de cohérence relevées dans notre corpus. Ils témoignent plutôt des trois catégories prévues par le schéma d'annotation d'UAM. Mais comme nous verrons dans la section 6.2.1, il y a un certain nombre d'autres phénomènes liés à la cohérence qui méritent d'être pris en compte.

6.1.1.3 « Pragmatic-error → erreurs de registre »

Les 152 éléments signalés ici ne tiennent compte que de trois phénomènes: (i) les items relevant d'expressions dites argotiques et qui ne conviennent pas par conséquent au genre textuel exigé des sujet-participants en contexte de rédaction institutionnelle, (ii) des contractions qui ne sont pas de manière générale acceptées dans le contexte des écrits universitaires et enfin (iii) certains choix lexicaux relevant d'un niveau de langue jugé trop familier ont également été annotés. Ces erreurs de registre sont donc tout à fait acceptables dans le système linguistique, mais sont jugés inacceptables dans certains contextes de rédaction.

10. **Cause* $Because$ on the other side of the coin, countries and companies […]. (txt_022_sm1)
11. There are **loads of* $several$ issues to address, starting with the fate of the eurozone. (txt_004_sm2)

A travers ces statistiques brutes, il devient clair que les erreurs pragmatiques sont pertinentes et ont, de ce fait, toute leur place dans l'analyse des productions écrites en anglais langue étrangère. De plus, le fait que nous identifions des erreurs de cohésion, de cohérence et de registre dans des énoncés qui – à première vue – sont grammaticalement corrects nous permet de réitérer notre postulat selon lequel la maitrise des règles syntaxiques d'une langue donnée ne coïncide pas avec une maîtrise pragmatique sous-jacente. Nous pensons que cette « maîtrise supplémentaire » survient souvent après l'acquisition d'un certain bagage linguistique dans la langue cible. Ce constat explique, entre autres, l'engouement autour des travaux (de plus en plus nombreux) portant sur les erreurs de cohésion et de cohérence tant en didactique de langue étrangère qu'en didactique de langue maternelle (cf. section 2.3.2). Nous reviendrons de manière approfondie sur un certain nombre de ces erreurs, notamment dans la section 6.2.1 qui introduit les erreurs d'acceptabilité textuelle.

6.1.2 Les erreurs de mise en phrase

Outre les erreurs liées aux questions de vocabulaire, aux structures syntaxiques sous-jacentes à la langue (grammaire), aux conventions institutionnalisées (genre, registre) ou à la maîtrise des quelques valeurs ou tournures sémantiques qui peuvent faire défaut aux apprenants de langue étrangère, il importe maintenant de porter notre attention sur les erreurs identifiées au niveau des structures proprement phrastiques et propositionnelles. C'est-à-dire, ces structures qui sont non seulement assujetties à la grammaire interne de la langue mais aussi tout singulièrement assujetties

au rapport phraséologique qu'entretiennent certains mots avec un autre mot ou certains types d'expressions, voire certaines tournures qui sont tout simplement « propres » à la langue. Nos appelons les erreurs liées à ce genre de phénomènes des erreurs de mise en phrase[97]. Celles-ci peuvent renvoyer à la fois aux unités lexicales complexes (comprendre les unités multi-mots) et aux unités propositionnelles plus large. De manière globale, on en dénombre un total de 687 occurrences : 393 au premier semestre et 294 pour le deuxième. Ce qui revient à une différence chiffrée de 99 d'un semestre à un autre ou à une baisse globale de vingt cinq pour cent.

les erreurs de mise en phrase	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
transferred-phrasing	45	61	106	16	35,56
other-phrasing-error	86	52	138	-34	-39,5
phraseology-error	83	28	111	-55	-66,3
phrase-segment-incomplete	178	152	330	-26	-14,6
response-formulation	1	1	2	0	0

Tableau 34 : Répartition des erreurs dites de « mise en phrase »

(vi) Erreur de transfert. (t*106 occurrences*). Le problème souligné ici est que certaines phrases identifiées dans notre corpus relèvent clairement d'une traduction mot à mot – et en dépit du fait que certaines peuvent ressembler à des expressions tout à fait correctes, d'autres (i) relèvent du « non-sens » voire (ii) créent un sens différent que celui voulu au départ.

12. **In the same order* $In a similar vein$, programs in which [a] company or country reduces or compensates for its carbon emissions […] (txt_006_sm1)
13. Would the economy collapse **in front of* $in the face of$ such a change? (txt_051_sm1)

(vii) Erreurs diverses de mise en phrase. (t*138 occurrences*). Dans cette catégorie, on relève des exemples de mise en phrase qui sont difficilement caractérisables, en raison de leur singularité et le fait qu'elles sont majoritairement toutes différentes les unes des autres. Ces cas de figure sont signalés dans cette dernière sous-catégorie où ils sont étiquetés « maladresse », faute d'un meilleur terme – dans la mesure où la mise en phrase en elle-même pose non seulement des problèmes de compréhension mais également d'acceptabilité en termes d'usage normé ou encore d'usage attendu.

[97] Ce terme renvoie tout simplement à ce que certains appellent des erreurs de phraséologie, il sera davantage précisé dans la section 8.2.2. Notons, par ailleurs, que le schéma d'annotation d'UAM avait uniquement prévu une seule « macro-catégorie » d'erreurs de mise en phrase (*phrasing*, en anglais), sans aucune sous-catégorie. Les sous-catégories ont été ajoutées, par nos soins, au fur et à mesure que nous avancions dans l'annotation : et ce, de manière à bien distinguer entre les différents types de phénomènes observés à ce niveau. Notons, par conséquent, à ce titre, que la catégorisation se veut principalement expérimentale. Nous en reviendrons – dans la section 8.4.2 – à cet aspect expérimental qui s'est vu intégré par la suite (par le concepteur du logiciel) dans la nouvelle version du schéma d'annotation.

(viii) Erreur de défigement phraséologique. (t*111 occurrences*) L'erreur apparait quand l'apprenant de langue étrangère procède par une « délexicalisation » d'une unité multi-mots figée dans la langue cible. Il peut s'agir d'un véritable « défigement » d'une unité fixe comme (i) dans certaines expressions idiomatiques ou certaines phrases prépositionnelles comme le suggère l'exemple (14), ou encore (ii) dans certaines unités multi-mots qui entretiennent une forte relation collocationnelle[98] les unes avec les autres.

14. Namely, it will not always be the case because this action can **in contrary* $on the contrary$ decrease consumers' consumption and [...] (txt_026_sm1)
15. It seems *now[sic]* clear that the **true problems* $real problem$ here is to know [...] (txt_026_sm1)

(ix) Le segment est incomplet. (t*330 occurrences*) L'erreur porte sur le fait que la mise en phrase n'est pas en adéquation avec le « sens voulu » par le sujet-participant. Il n'est pas question ici de deviner le sens voulu par l'apprenant, mais plutôt de constater que ce qui a été rédigé n'est pas complet sémantiquement – et ne permet, de ce fait, pas à la structure propositionnelle d'exister en tant qu'unité de sens à part entière.

16. Companies such as multinational firms, **based on a competitivity law* $which are regulated by competition law$, are not ready to lose competitiveness for going greener [...] (txt_042_sm2)
17. We can add that, after the crisis of 1929, the wealthy countries **start to be indebt* $indebted themselves trying$ to spur the economic activity with important *public spending[sic]*. (txt_048_sm1)

En bref, le fait que nous avons dû ajouter les quatre sous-catégories d'erreurs de mise en phrase à la taxonomie initiale d'UAM suggère (i) que l'étiquetage initial n'était pas suffisant et (ii) par voie de conséquence, qu'il y a un certain nombre de différences (comme nous l'avons vu à travers les exemples (*vi)* à (*ix),* et respectivement *(12)* à *(17)*) dans les occurrences erronées identifiées à ce niveau de production en langue étrangère. Pour toutes ces raisons, l'ensemble des quatre catégories présentées ici sera discutée davantage dans notre deuxième volet d'annotation.

[98] Ce point est amplement discuté dans la deuxième partie de ce chapitre, notamment dans la section 6.2.1, à propos de deux types d'erreurs d'acceptabilité textuelle (cf. erreurs sémantiques et erreurs de mise en phrase).

6.1.3 Les erreurs de connecteur

(t*104 occurrences*) On signale notamment ici (i) l'absence d'un connecteur ou mot de liaison approprié voire (ii) un choix lexical à la fois structurellement et sémantiquement inadapté au contexte textuel. Les deux cas peuvent provoquer une rupture thématique dans la mesure où la distribution informationnelle se trouve soudainement discontinue. Dans le premier cas, le choix n'est pas en adéquation avec la progression thématique tandis que dans le deuxième cas, il y a un changement brusque dans l'argumentation qui aurait gagné à être signalé de manière à faciliter la transition d'un argument à un autre. Toutefois étant donné que ces erreurs dépendent du contexte textuel, il faut disposer des empans textuels conséquents – en termes de pré et post-textes - pour apprécier les erreurs à leur juste valeur.

18. Globalisation can be *economicaly[sic]* defined as the development of international exchanges that led countries to be more *interdependant[sic]*. This *processus[sic]* started a long time ago but really *growed[sic]* after the Second World War, thanks to the new industries and technologies, like *internet[sic]* which helped to spread information and **so* $in a manner of speaking$ globalisation. (txt_047_sm1)
19. First of all, this question of better job opportunities can be asked because of the increase of women rights in terms of education this past century. Mentalities have dramatically evolved, especially when we remember that fifty years ago women were *considerate[sic]* less smart than men, because of their biology, and **though* $for the reason that$ they could not *reach[sic]* high levels of education. (txt_042_sm2)

Bien que la conjonction de coordination « so » dans l'exemple (18) ait le sens de « therefore » ou permette – de manière générale – d'introduire une remarque finale (comme l'apprenant a visiblement tenté de le faire), son usage ne convient pas à ce contexte textuel précis. Sur le plan phraséologique « *spread globalisation » constitue une sorte de collocation impropre, tandis que sur le plan de la structuration informationnelle, le lien entre l'argument principal (en conséquence la proposition principale) et le mot « globalisation » ne peut pas être clairement établi. D'où l'emploi inapproprié de ce connecteur. L'exemple (19) est d'une certaine manière similaire au précédent. Mais on pourrait soulever deux problèmes supplémentaires : un au niveau sémantique et l'autre au niveau du registre[99]. Ces deux exemples d'erreurs de connecteur (18 et 19) font, de ce fait, écho à l'ensemble des erreurs relevées dans cette catégorie. Et comme nous l'avons vu dans d'autres

[99] Il convient de préciser que l'usage de « though » est ostensiblement plus fréquent dans des échanges informels, tandis que l'usage de « although » est considéré plus approprié dans un écrit formel.

catégories d'erreurs – qui ont été mises en avant notamment dans les deux sections précédentes – la caractérisation de certaines erreurs n'est pas toujours clairement imputable à un seul trait distinctif : c'est-à-dire, il n'est pas toujours possible de soutenir que telle ou telle erreur de connecteur est strictement textuelle ou relève strictement du système linguistique. Nous allons examinons certains de ces cas épineux dans la sous-section suivante.

6.1.4 Les chevauchements entre système et texte

Après un examen minutieux des erreurs présentées dans les trois sous-sections précédentes, il nous parait judicieux de souligner que la délimitation (ou la frontière) entre erreurs textuelles et erreurs du système est loin d'être évidente. Dans certains cas, si nous ne prenons en compte que la nature ou la fonction d'un item (au dépens du sens), il devient clair que la grammaticalité de l'item ne peut pas être remise en cause – comme dans l'exemple (23) et dans la proposition principale de l'exemple (25) ci-dessous. Dans ces cas, uniquement le sens de l'ensemble textuel permet d'apprécier le caractère inacceptable de l'occurrence. Dans d'autres cas, nous constatons que la grammaticalité aussi bien que le sens résultant de l'ensemble créent un effet « d'impropriété de langage ».

20. So the countries have to reduce **his* $their$ carbon emissions [...]. (txt_088_sm2)
21. For example more and more airlines propose to **its* $their$ passengers [...]. (txt_092_sm1)
22. Anyone in **its* his right mind would say the Union is doomed […]. (txt_004_sm2)
23. **During the center* $middle$ *of the twentyth[sic] century,* [...]. (txt_055_sm2)
24. To some extent, the Telegraph's statement is right because some areas are less likely to be studied **on degrees* $at the university level$. (txt_057_sm2)
25. This idea concerns the public sector too, which has adopted the new Public Management [...]. Although the working world **become* $has become$ a pursuit of benefits, some trainings **stayed* $have remained$ a reference on the labour Market. (txt_08_sm2)

Dans les exemples (20), (21) et (22), nul ne saurait réfuter le caractère agrammatical des énoncés – provoqué notamment par l'incompatibilité des pronoms : les deux premiers sont erronés aussi bien en termes de type de pronom qu'en termes d'accord en nombre, tandis que l'exemple (22) s'accorde bien en nombre mais n'est pas du type pronominal attendu dans ce contexte précis. Mais au-delà de la grammaire, on a l'impression que le texte restreint aussi le choix des pronoms et joue donc un rôle dans ces erreurs. Cependant, l'influence du texte est plus facilement identifiable dans les trois autres exemples. Dans l'exemple (23), la phraséologie est le principal élément en cause :

que ce soit ‘during the center’ ou ‘the center of the century’, les constituants sont syntaxiquement[100] – voire dans une certaine mesure sémantiquement – acceptables. Mais en termes d’ensemble phraséologique, l’assemblage pose problème. L’exemple (24), comme le précédent, partage un caractère agrammatical et un caractère inacceptable sur le plan phraséologique. De plus « *become » dans l’exemple (25) est tout d’abord agrammatical puisque l’on remarque le problème d’accord entre le sujet grammatical de la phrase et le groupe verbal. Mais au-delà de cet écart de syntaxe, le contexte textuel exige l’emploi de l’aspect perfectif sur les deux verbes présents dans l’énoncé. L’erreur portant sur la concordance des temps est donc d’ordre textuel.

En définitive, la difficulté que l’on soulève dans l’identification d’une erreur comme étant proprement linguistique ou proprement textuelle revêt un caractère particulièrement épineux. A travers les six exemples ci-dessus, sans oublier bien entendu ceux des trois sections précédentes, il convient d’admettre que la distinction entre nos deux grands types d’erreurs n’est pas toujours aussi clairement délimitable que l’on aurait espéré. A la lumière de ces constats, passons au deuxième volet d’annotation (section 6.2) dans lequel nous nous intéressons davantage aux erreurs qui sont manifestement (et uniquement, pour ainsi dire) d’ordre textuel.

6.2 Les erreurs textuelles (volet 2)

Rappelons tout d’abord que seules les occurrences dépourvues de tout agrammaticalité ont été examinées dans ce deuxième volet d’annotation. De plus, contrairement au schéma d'annotation d’erreurs intégré à l'UAM CorpusTool et aux schémas issus des trois métafonctions de la linguistique systémique fonctionnelle, les annotations dans ce volet ont été développées au fur et à mesure que les erreurs textuelles ont été relevées.

6.2.1 La catégorisation des erreurs d’acceptabilité textuelle

Étant donné l'aspect pilote de cette annotation, nous avons préféré des catégories générales à des catégories d’une granularité fine (*fine-grained*, en anglais) dans le but d’éviter autant que possible les chevauchements éventuels de catégories trop précises. Le tableau suivant donne un aperçu général des statistiques brutes et des différentes erreurs observées.

[100] Reconnaissons que l’on peut également soutenir que l’exemple est grammaticalement inacceptable, dans la mesure où la préposition « during » n’accepte d’être suivie que par un groupe nominal indiquant une période temporelle et non une période ‘spatiale’ (en termes de distance). L’erreur viendrait alors de l’ambigüité polysémique de « middle » qui est sémantiquement plus « spatial », mais peut renvoyer à des périodes temporelles principalement dans des expressions semi-figées du type « in the middle of the night », etc.

Acceptabilité-type	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
E. référentielle	47	24	71	-23	-48,93
E. sémantique	85	75	160	-10	-11,76
E. de cadrage	19	17	36	-2	-10,52
E. de coordination	24	31	55	7	29,16
E. d'agencement	26	24	50	-2	-7,69
E. de focus	65	72	137	7	10,76
E. de mise en phrase	73	102	175	29	39,72
Total	339	345	684	6	1,76

Tableau 35 : Catégories d'erreurs textuelles

Comme le démontre le tableau 35, sept types d'erreurs ont pu être identifiés à ce niveau – avec un total de 684 items étiquetés. Soulignons, toutefois, que ce chiffre représente les items obtenus dans la moitié du corpus, après l'étape de rééquilibrage - c'est-à-dire, après avoir sélectionné les énoncés dépourvus de toute agrammaticalité[101].

(i) Erreur référentielle. (t*71 occurrences*). Ces erreurs font écho à celles identifiées en section 6.1.3 (cf. erreurs pragmatiques, et plus précisément erreurs de cohésion), comme étant des problèmes de saillance et d'incompatibilité entre référents. La différence principale réside dans le fait que ces dernières ont souvent conduit à des phrases agrammaticales tandis que les phrases dans ce deuxième volet d'annotation demeurent grammaticalement acceptables. Nous avons donc affaire ici à des erreurs de référence anaphorique avec lesquelles il s'avère impossible d'établir un lien clair entre le substantif et le pronom auquel il est supposé se référer.

26. The educational system has increased the number of *high degrees[sic]* but it doesn't mean that people are able to work in firms after their degrees. The 2012 Davos summit *give[sic]* us an answer to **this problem* $##$[102]. (txt_006_sm2)
27. It is much more difficult *where[sic]* you are the President's wife but some of them have succeeded in *outcoming[sic]* and outranking **them* $##$. (txt_024_sm2)

En effet, hormis le problème d'accord en nombre et le problème d'incompatibilité entre référents pronominaux que nous avons mis en avant dans la section 6.1.1., l'erreur référentielle conduit irrémédiablement ici à un problème d'ordre logico-sémantique. La construction, à proprement parler, n'est pas affectée mais la compréhension du message complet n'est pas non plus

[101] Cf. la section 4.2.4 pour un rappel sur la sous-division du corpus.
[102] Pour rappel (cf. La liste des sigles, abréviations et conventions), « $##$ » signifie qu'aucune réponse (i) ne peut être proposée ou (ii) ne peut être inférée au vu de l'ambigüité de la phrase. L'idée est simplement de ne pas à apporter des informations supplémentaires « à la place » de l'apprenant-scripteur.

« totalement assurée ». Autrement dit, l'erreur référentielle freine le lecteur au niveau de la fluidité et de la clarté de l'information. Le lecteur se voit donc contraint d'établir voire d'induire, malgré lui, les rapports référentiels (qui sont insuffisamment saillants) entre les différents constituants du texte.

28. They will build a European Union but despite **this* $##$ they make their own policy and refuse to have a common policy. **That* $##$ is the same situation in terms of [...]. During the Arabic Revolution *thousand[sic]* of illegal migrants have entered *in[sic]* Italy particularly in Lampedusa and the Italian government coped with this problem without the European Union while the majority of these migrants would join other European Countries… In term of monetary *issus[sic] *that* $##$ is the same case. (txt_093_sm2)
29. Advertising was created to *tempte[sic]* people, to make them *consum[sic]* more and more. It is the biggest tool for consumption. So if it disappear[sic] this industry would *loose[sic]* a lot. People would probably be less tempted and would buy less. Advertising is here to make us believe that we need **this new product* $##$. Even if *Ø[sic]* lived without it *during[sic]* twenty years. **It* $##$ is what happened with the Iphone or other *smartphone[sic]* +$##$, we are now wondering, how would we lived without it. And this idea of new need was created by an ad. So without it, it would probably be difficult to convince people that they need **this product* $##$. (txt_121_sm1)

De plus, comme en témoignent les exemples (28) et (29), quand les erreurs référentielles sont multiples et se suivent les nombreuses possibilités d'interprétation conduisent à un effet déconcertant – ce qui ne facilite ni la lecture et ni compréhension du texte selon le sens ou le positionnement précis voulu par l'apprenant-scripteur.

(ii) Erreur sémantique. (t*160 occurrences*). Il arrive aussi que les erreurs soient d'ordre purement sémantique : c'est-à-dire que la phrase soit grammaticalement correcte mais conduit à un non-sens, sur le plan sémantique. Ce type d'erreur a été repéré sur l'ensemble des classes grammaticales et a priori semble être plus que le simple effet d'un « transfert de sens » d'un terme de la langue française vers la langue anglaise. Cela dit, l'apprenant respecte la morphosyntaxe de la langue anglaise dans l'utilisation de son emprunt.

30. Even if people are conscious of this inequality and mentalities **are moving* $are changing$, men are still superior to women in the labour market. (txt_067_sm2)

31. During that time, women's move[ment] gained ground very fast and women asked for more equality between sexes in general. Today educational parity **has settled* $has become of age$ and we can even talk about women's superiority [...] (txt_071_sm2)
32. And we will try to present different measures in order to find **issues to* $solutions for$ this problem. (txt_115_sm2)

En effet, comme nous le verrons plus en détail dans la section 8.2.2 et tout particulièrement dans les sections 8.1 et 8.2.2.1, les erreurs sémantiques sont beaucoup plus importantes en nombre que prévues : elles comptabilisent 23% des occurrences annotées dans l'ensemble des erreurs d'acceptabilité textuelle. Les exemples ci-dessus mettent en avant les trois types d'items les plus fréquents que nous avons relevés dans cette catégorie. L'exemple (30) fait état d'un choix erroné dont on peut commodément « deviner » le sens voulu par l'apprenant-scripteur. **Move* et *$change$* partagent une certaine isotopie sémantique – le premier peut se définir globalement comme un changement de position d'un endroit à un autre, tandis que le deuxième traduit le fait de subir des modifications. On pourrait en conclure que c'est le sens précis de ces termes qui fait défaut à l'apprenant.

Les exemples (31) et (32) ne sont pas du même ordre que le précédent et renvoient dans une certaine mesure à des erreurs de mise en phrase voire plus précisément à des erreurs de phraséologie lexicale (cf. point (*iv*), ci-dessous). Toutefois, étant donné que le sens premier des items annotés est totalement inadapté (i) au besoin informationnel et (ii) au contexte textuel, affirmer que les erreurs sont plus phraséologiques que sémantiques suppose en quelque sorte privilégier une seule explication sur l'origine de ces erreurs alors que plusieurs sont possibles. A titre d'illustration, « *has settled » dans (31) peut signifier (i) que la parité à l'école ne constitue pas un problème en soi aujourd'hui (et donc que la question est tout simplement réglée) – ou si on se positionne par rapport à l'ensemble des arguments avancés dans le texte n° 072_sm2 – (ii) que la parité est désormais bien ancrée dans la société (ce qui suppose qu'elle n'a pas toujours été le cas et que le cheminement jusqu'à aujourd'hui a été long). Selon « l'interprétation » que l'on attribue à l'erreur sémantique « *has settled », on pourrait la corriger par « $is now accepted$ ou $has come of age$ – ce qui ne traduit pas la même réalité. En somme, les erreurs strictement sémantiques témoignent, à notre sens, d'une sorte de connaissance lexicale superficielle, c'est-à-dire qu'il y aurait chez l'apprenant une méconnaissance des propriétés sémantiques inhérentes.

(iii) Erreur de cadrage. (t*36 occurrences*). Les erreurs relevées ici affectent tout particulièrement les phrases qui les suivent voire l'ensemble du paragraphe qui les succède. En effet, la

fonction principale de l'élément sur lequel porte l'erreur est de fournir un angle d'interprétation ou un angle de délimitation des informations qui sont présentées. L'erreur survient alors quand l'élément ou le cadre ne remplit pas correctement sa fonction : notamment en raison du fait (i) qu'il n'est pas approprié au contexte ou (ii) qu'il résulte d'un usage sémantique erroné. Rappelons par ailleurs qu'un contexte textuel large est un facteur décisif sans l'appréciation des items annotés ici.

33. **In the last but one sentence* $##$, Mr Gül underlines the fact that this interdependence, created by the process of Globalisation, is responsible for *Ø[sic]* global economic crisis, such as the one we are facing today which is by the way one of the most devastating *crisis[sic]* the world has even been confronted to. This statement is true in the *facts[sic]* as the current crisis was triggered in America particularly because of the subprime mortgage crisis of 2008. (txt_079_sm1)
34. People are becoming more and more educated. Whereas in the XIXth century, education was only *reachable[sic]* by *higher sphere[sic],* we notice that since the second world war, the number of students is increasing each year. **In front of this evolving process* $##$, the world of work has also deeply changed, but often not in the same way: unemployment *is[sic]* skyrocketing for the last decades in a lot of countries. So we can wonder if the Educational system is really adapted to the way the world of work changes. (txt_076_sm2)
35. **Since its first day* $From day one$, the European Union has been growing step by step. Despite of a lot of difficulties *through years[sic]*, the EU still *exist[sic]* and *develop[sic]* itself now, thanks to so many political leaders. (txt_83_sm2)

Si l'on examine chacun des trois exemples ci-dessus, l'erreur se trouve systématiquement sur ce que Charolles (2009) appelle un cadratif : à savoir ce qui permet d'indexer une partie du discours à venir. Charolles souligne le fait que certains adverbiaux « sont susceptibles, quand ils sont employés à l'initiale de phrase, de jouer un rôle dans l'organisation du discours du fait qu'ils peuvent porter non seulement sur le contenu de leur phrase d'accueil mais aussi sur celui d'une ou plusieurs autres apparaissant dans la suite ». Si l'on adopte cette position, il devient évident que la portée du cadratif s'étend distinctement – dans l'exemple (33) – sur les phrases « Mr Gül underlines […] » et « This statement is true […] ». Le fait donc d'avoir un cadratif dont le sens demeure incertain ou indéterminé constitue un frein majeur à la (bonne) lecture et la (bonne) compréhension du message.

Dans les exemples (34) et (35), l'erreur porte à nouveau sur un cadratif. Mais nous relevons cette fois-ci un certain nombre de chevauchements indéniables qui méritent d'être explicités. D'abord, le sens du cadratif peut dans les deux exemples être facilement inféré. Ce qui semble constituer un problème supplémentaire réside dans la phraséologie du segment cadratif. A titre d'illustration, dans l'exemple (34), « *In front of » et « *this evolving process » renvoient de manière distincte à deux erreurs dont l'usage lui-même fait naître un certain effet d'impropriété de langage. Mis à part cet effet, l'emploi (i) de « *In front of » et « *this evolving process » ensemble et (ii) à l'initiale de la phrase donne une autre dimension erronée. Et comme nous pouvons le voir, les phrases qui succèdent au cadratif sont en quelque sorte l'explicitation de celui-ci – ou du moins, elles l'emploient comme un élément d'ancrage central pour l'argumentation qui est développée par la suite.

(iv) Erreur de coordination. (t*55 occurrences*). Il s'agit ici d'erreurs portant sur l'absence de conjonctions de coordination : à savoir, le lien (la relation syntaxique) entre certains items n'est pas suffisamment saillant – ce qui influe sur l'intelligibilité et l'acceptabilité de l'énoncé et ses contours immédiats. Plus précisément, le problème est principalement d'ordre paratactique : au lieu d'une structure asyndétique, l'apprenant aurait gagné à utiliser une structure syndétique pour mieux coordonner et ainsi expliciter la relation existante entre les différents items présentés.

36. In other words, we chose economic growth over the environment. But we try to compensate with **programs, measures* $(different?) programs and measures$. Are they really effective? (txt_063_sm1)
37. Some *environmentalist[sic]* organizations check the carbon offsetting of companies and deliver a label in order to fight against greenwashing. For example Rainforests alliance Lipton tea maps in Kenya or various parameters: **carbon offsettings, decent working conditions, environmental respect* $including for example carbon offsetting, decent working conditions and environmental respect$. (txt_092_sm1)
38. The point is the society has changed in favour of women, but **the educational parity, the government parity,* $the educational parity and the governmental(?) parity$ will not be enough. (txt_100_sm2)

Dans ces exemples, on constate une juxtaposition de plusieurs éléments sans coordination ou subordination entre eux. La responsabilité incombe au lecteur d'établir les rapports syntaxiques manquants. De manière générale, il suffit d'ajouter une conjonction de coordination pour corriger

l'erreur, comme dans les exemples (36) et (38). Par contre, dans l'exemple (37), on a besoin de préciser le type de relation syntaxique entretenue entre la proposition principale (ou la proposition précédente, dans ce cas précis) et les éléments qui sont énumérés. En bref, soulignons que la frontière entre ce que nous appelons les erreurs de coordination n'est pas aussi distinctement délimitable que l'on aurait espéré. Certaines erreurs se corrigent assez facilement (comme dans 36 et 38), tandis que d'autres se rapprochant à des erreurs d'ostension (cf. les différents types d'erreurs de focus ci-dessous, et la section 8.2.2.2 pour une discussion approfondie) sont plus « délicates » à corriger.

(v) Erreur d'agencement. (t*50 occurrences*). Il est question ici d'erreurs d'agencement portant principalement sur les phrases adverbiales. Les occurrences annotées sont sensiblement similaires à certaines erreurs[103] du chapitre précédent, mis à part, bien entendu, l'effet d'agrammaticalité provoqué dans ce dernier. Cela laisse penser par conséquent que ce problème est profondément ancré dans l'interlangue de l'apprenant (du moins, plus qu'on ne le pensait), notamment en raison du fait qu'il réapparait même après que l'apprenant ait vraisemblablement acquis une certaine maturité syntaxique et lexicale.

39. Simone de Beauvoir argues that women **always* have $always$ been considered as the "second sex" [...] (txt_066_sm2)
40. Nowadays, globalisation is a massive and undeniable phenomenon affecting **more and more* our way of life $more and more$. For example, sixty years ago, people wearing jeans, baskets and t-shirts were rare. (txt_112_sm1)
41. We need to question our ways of life in order to pass into a low-carbon society because this shift *holds off[sic*] more opportunities than constraints. Indeed **already* the countries that have $already$ embraced this change recognise that it can bring jobs, economic growth and a better quality life. (txt_092_sm1)

Ce qui mérite d'être signalé avant tout ici est la variabilité des erreurs d'agencement. Certaines sont placées de façon erronée (i) devant le verbe (exemple 39) ; (ii) directement après le verbe (exemple 40) ; ou (iii) ou en position initiale de la phrase (exemple 41). En prenant donc en compte ces variations, sans oublier, les différents types d'erreur portant sur l'agencement syntaxique mis en avant dans le chapitre V, nous pouvons soutenir que l'emploi des éléments non-obligatoires syntaxiquement (c'est-à-dire, en dehors du sujet grammatical, le verbe et son complément)

[103] Pour rappel, les erreurs d'agencement (ou les erreurs d'ordre ou de position syntaxique) du premier volet d'annotation portaient sur cinq catégories distinctes : à savoir les ajouts ('adjunt'), les déterminants, les pré-modifieurs, les post-modifieurs et les marqueurs de négation ('do not'). Cf. section 5.1.2.

constitue un problème qui nécessite une attention particulière chez nos apprenants-scripteurs. Ce problème traduit une méconnaissance, non seulement lexicale ou proprement sémantique, mais tout simplement au niveau de l'emploi syntaxique privilégié de ces « ajouts ». Nous discuterons davantage de ces problèmes d'ajouts, en mettant l'accent notamment sur les différents types d'ajouts problématiques et leur fréquence d'erreurs dans la section 6.2.2, ci-dessous.

(vi) Erreur de mise en phrase. (t*175 occurrences*). Notons qu'il y a un rapprochement à faire entre les différents types d'erreurs signalés en section 6.1.2 – qui ont majoritairement provoqué des agrammaticalités – et les erreurs présentées dans cette section. Ces dernières portent sur des empans textuels dont la structure grammaticale est jugée correcte mais dont le sens résultant de l'assemblage pose des questions d'acceptabilité. En effet, sans passer en revue l'ensemble de ces dernières, on pourrait tout simplement signaler qu'elles se distinguent de plusieurs manières mais que nous les regroupons en deux grands groupes : à savoir les erreurs de transferts phraséologiques et les erreurs de fabrication phraséologique.

(A) Pour le premier type, il s'agit en principe (i) des erreurs portant sur des locutions fixes (par exemple, les expressions idiomatiques) que les apprenants ont transposé de façon littérale d'une langue déjà connue à la langue cible ; ou (ii) des erreurs portant sur des locutions semi-figées du type (a) « lexical bundles » au sens de Biber & Conrad (1999), (b) collocations et (c) colligations qui sont également transposées d'une autre langue vers la langue cible.

(B) Pour le deuxième type, on est généralement dans un cas de figure de création, d'utilisation ou de modification des unités multi-mots (ou des unités composés) existant en langue anglaise. L'erreur vient en principe de la non-application d'une règle d'usage propre à la langue cible : par exemple, une contrainte syntaxique qui nécessite une inversion des mots. A titre d'exemple, dans des phrases conditionnelles dans lesquelles « if » est omis, '*Were it not for the harsh tax laws, she would not have emigrated'*, si l'apprenant oublie l'inversion le résultat est quelque peu biscornu.

Qu'il s'agisse du premier ou deuxième type, l'apprenant n'a tout simplement pas connaissance (ou n'applique pas ses connaissances) du fait qu'un terme utilisé entretient des rapports privilégiés avec un autre. Le résultant phraséologique peut, de ce fait, conduire à des non-sens ou un sens différent de celui visé par l'apprenant, et par conséquent à un sens qui ne

convient pas à la fluidité de l'information. Deux exemples sont fournis illustrant principalement le premier type : c'est-à-dire, le (A) (i).

42. Quotas are also **at the order of the day* $a topical issue/(currently) on the agenda$ [...]. (txt_066_sm2)
43. The evolution of women's right has been **taking a huge place* $ playing a significant role$ in our society for several decades, and it is nowadays one of the main subjects of controversy. (txt_070_sm2)

(vii) Erreur de « focus ». (t*137 occurrences*). Ces erreurs surviennent dans une grande mesure dans la périphérie droite de la phrase, voire même en tout dernière position. Etant donné que plusieurs traits distinctifs ont été repérés, cette catégorie a été subdivisée en trois parties et des exemples sont fournis ci-après.

focus-type	**sm1**	**sm2**	n***(sm1+sm2)***	***d(sm2-sm1)***	***(%)***
Ostension	18	17	35	-1	-5,55
Progression thématique	14	8	22	-6	-42,85
Progression temporelle	33	47	80	14	42,42
total	65	72	137	7	10,76

Tableau 36 : zoom sur les erreurs de type "focus"

a) Erreur d'ostension. (t*35 occurrences*). Il s'agit ici d'erreurs portant sur le segment phrastique dans lequel on cherche à fournir des exemples ou des points d'éclaircissement. Dans ces cas précis, la relation hypotactique (à savoir, la relation de dépendance ou de subordination) n'est pas clairement établie. L'erreur d'ostension survient alors quand la responsabilité revient au lecteur potentiel de créer un rapprochement aléatoire entre l'illustration et le segment textuel auquel il est rattaché. Que cette erreur se manifeste dans un segment textuel court ou un segment textuel long, la continuité de la chaîne informationnelle se trouve souvent interrompue au vu de l'inintelligibilité du message.

44. Some studies talk about a lack of ambition for women, **a link of self-censure* $##$. (txt_066_sm2)

b) Erreur de progression thématique. (t*22 occurrences*). Hormis le problème posé par l'erreur d'ostension, lorsqu'aucun lien n'est clairement établi entre des séquences textuelles ou des phrases qui se suivent, on a affaire à une erreur de progression thématique[104] : c'est-à-dire une

[104] Etant donné la multiplicité de définitions existant dans la littérature, (cf. Combettes & Tomassone 1988 ; Carter-Thomas 1999a, 1999c ; Fontaine & Kodratoff 2002), la nôtre se veut assez générale en renvoyant à toute rupture

rupture soudaine dans la chaîne informationnelle. Notons de plus qu'une erreur de progression thématique peut dépasser la phrase simple et provoquer ce que l'on appelle ordinairement un hors-sujet. Il est important de préciser donc, de la même manière que l'ensemble des erreurs annotées dans ce volet et tout particulièrement les erreurs de cadrage, que les erreurs de progression thématique nécessitent souvent un empan textuel assez large pour apprécier l'erreur à sa juste valeur. Toutefois un court segment singulièrement représentatif est présenté ci-après.

45. Unemployment is a crucial issue today. The rate of unemployment is particularly high for young people in countries such as Spain for instance. **Education is the key to the world of work* $##$. (txt_112_sm2)

c) Erreur de progression temporelle. (t*80 occurrences*). Cette erreur peut également être comparée à une sorte de rupture informationnelle mais dans une moindre mesure. En effet, l'erreur ne porte pas sur le cœur du message mais sur la prédication ou sur l'élément qui permet d'actualiser ou de tisser le fil de l'information. L'erreur à ce niveau n'est ni d'ordre sémantique ni d'ordre lexical, mais à mi-chemin à proprement parler entre le textuel et le grammatical. Ceci s'explique par le fait que la continuité textuelle exige dans un texte donné l'accord du temps grammatical. Et c'est justement cette condition qui n'est pas satisfaite ici.

46. A woman had to take care of her child, and she wasn't able to work. Working was a *man skill[sic]*. The point is the society has changed in favour of women, but the *educational parity, the government parity [sic] *will not be* $are not$ enough. The modern woman, who wants a dual-earner family, who's graduated and *independant[sic]*, will have her place when the entire society **will change* $changes$.

Dans l'ensemble, les sept types[105] d'erreurs d'acceptabilité textuelle présentés dans cette section renvoient à ceux qui ont été jugés les plus significatifs, en raison de leur fréquence statistique globale. Et bien qu'ils n'aient pas la prétention de représenter l'ensemble des erreurs d'acceptabilité textuelle existantes, ils ont tout de même permis de mettre en avant de véritables problèmes qui sont souvent mis à l'écart dans la classe de langue étrangère - dans l'espoir sans doute que la variable temporelle « fasse le nécessaire » pour réduire ces erreurs sans nom. Nous supposons, de

---

inattendue dans la chaîne informationnelle – notamment quand celle-ci semble détachée à la fois de ce qui la précède et de ce qui suit.

[105] Dans le présent travail, les trois sous-types d'erreurs de focus sont considérés comme un ensemble. Mais nous pensons, cependant, que des études complémentaires (avec un corpus d'apprenants d'un niveau en L2 anglais égal ou supérieur à B2) permettront de leur donner une validité propre (et donc de les « affranchir » de la catégorie d'erreurs de focus), notamment pour ce qui est des erreurs d'ostension.

plus, que cette mise à l'écart se produit en général parce que les enseignants n'identifient pas distinctement ces différents phénomènes (peut-être ne voient-ils pas l'intérêt) et donc ils ne posent pas, en tant que telles, ces questions épineuses dans l'évaluation des productions écrites.

D'un autre coté, les étudiants, à leur tour, sont incapables de comprendre pourquoi certains éléments ont été signalés comme étant erronés (peut-être par un simple soulignement ou un point d'interrogation) et sont, de ce fait, incapables de procéder sciemment à une autocorrection. L'avantage donc, à notre sens, du signalement de ces erreurs d'acceptabilité textuelle est par conséquent double : tout d'abord, il est d'ordre pédagogique, car nous pensons que l'identification précise de tout élément qui s'avère problématique pour un étudiant ne peut que l'aider à prendre conscience du problème et à y remédier consciemment ; ensuite, il est d'ordre linguistique, car ce signalement apporte un argument de plus à tous ceux qui affirment que la maturité textuelle n'est pas le résultant direct d'une simple maturité syntaxique.

6.2.2 Le schéma expérientiel appliqué aux erreurs d'acceptabilité textuelle

Passons maintenant aux résultats de la ré-annotation des erreurs d'acceptabilité textuelle, selon le premier schéma issu de la métafonction systémique. Précisons tout d'abord à titre d'information que certaines erreurs annotées en section 6.2.1 ne sont pas prises en compte en section 6.2.2, 6.2.3 ou 6.2.4 puisqu'elles dépassent les frontières d'une ou plusieurs micro-unités individuelles. Par exemple, certaines erreurs de cadrage ou de mise en phrase ne peuvent être ré-annotées avec le schéma expérientiel puisqu'elles ne sont pas confinées à un seul des trois rôles expérientiels de base : autrement dit, elles peuvent concerner deux ou les trois rôles expérientiels en même temps, à savoir le procès, les participants ou les circonstances. Cela étant, le fait d'utiliser le schéma expérientiel pour ré-annoter les autres catégories d'erreurs d'acceptabilité textuelle s'est tout de même avéré, une fois de plus, très concluant : la figure 34 ci-dessous en témoigne à travers une répartition claire et sans appel.

En effet, parmi les trois rôles de base existant à ce niveau, les rôles circonstanciels se révèlent être les plus problématiques pour nos sujets-participants, avec 50% des occurrences comptabilisées. Ce résultat n'est pas une surprise en soi, étant donné que ce problème avait déjà été signalé comme étant particulièrement délicat dans le premier volet de l'étude. A ce résultat s'ajoutent de façon décroissante les microphénomènes portant sur des participants (38,31%) et des procès (11,21%). Notons que la répartition des erreurs a été complètement inversée par rapport au premier volet, laissant apparaître les problèmes qui sont profondément enracinés dans l'interlangue de l'apprenant

et qui mériteraient davantage d'attention si l'on souhaite éviter une certaine fossilisation de ces erreurs chez les apprenants les plus avancés.

Examinons maintenant ces résultats en détail. Parmi les items retenus comme erreurs circonstancielles, la grande majorité relève de celles ayant une fonction de projection – et plus précisément, celles qui permettent de donner un angle ou une « perspective nouvelle » au procès auquel la circonstance est liée. Selon Halliday & Matthiessen (2004 : 263) « l'angle circonstanciel » permet d'établir deux choses : (i) le *viewpoint* - au sens de « *to, in the view/opinion of, from the standpoint of* » et (ii) la source du sens de « *according to, in the words of* », et ainsi de suite. Il est important de souligner également que le problème de perspective constitue à lui seul un tiers des erreurs annotées sur l'ensemble des huit sous-catégories possibles.

47. The bankruptcy of Lehman's Brothers in 2008 $greatly$ affected **for an important part* the [E]uropean banks.(txt_098_sm1)
48. **On his mind* $In his opinion$, all countries are integrated in a world economy. (txt_107_sm1)

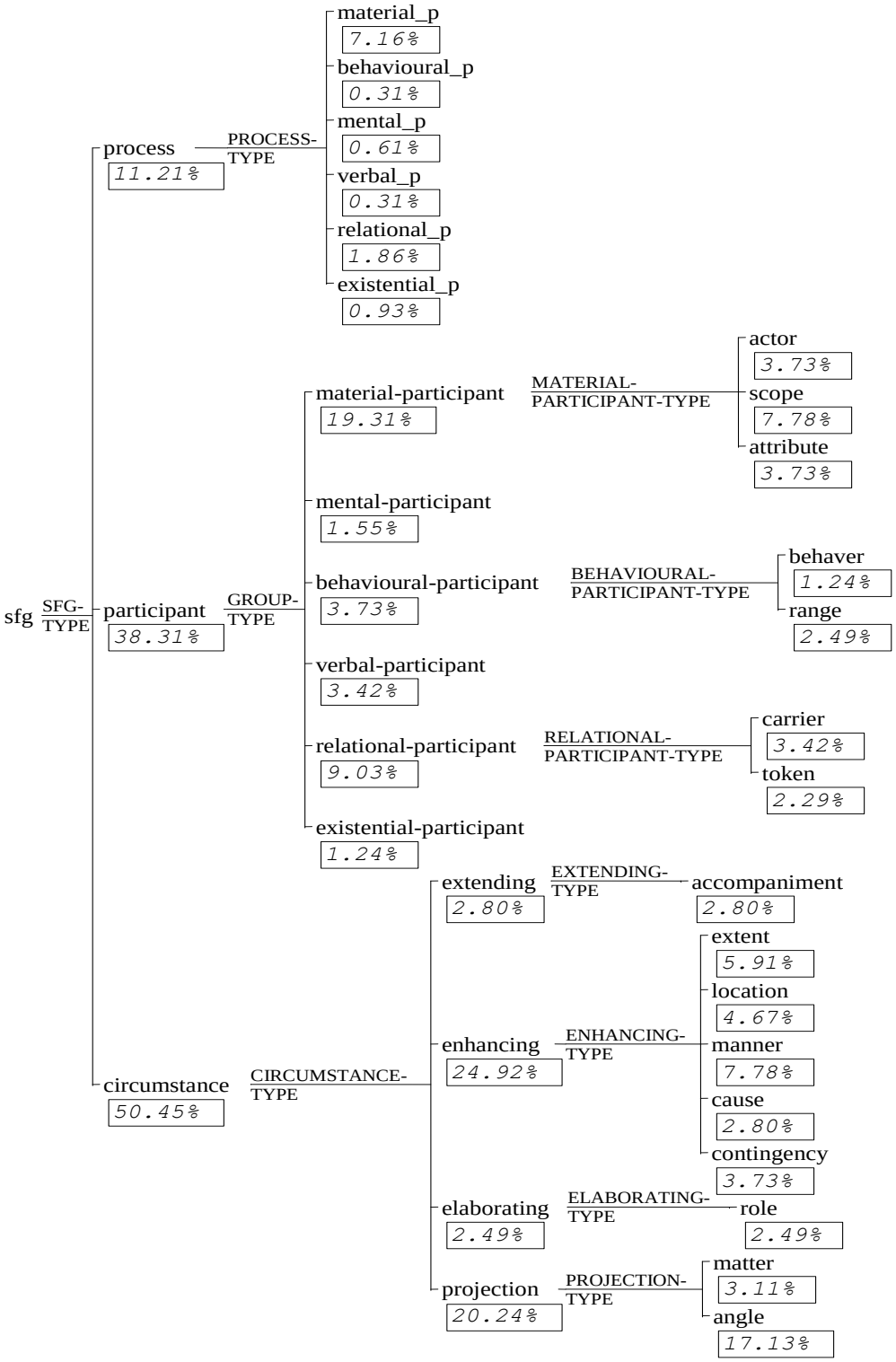


Figure 34 : Le schéma expérientiel au service des erreurs textuelles[106]

[106] Etant donné le nombre considérable d'étiquetages possible dans le schéma expérientiel, seuls les items les plus fréquents ont été retenus dans le troisième niveau portant sur les participants – de façon à éviter une illustration à rallonge avec des items peu pertinents. Les fréquences statistiques de 0 à 1 ont donc été omises.

S'ensuivent alors les trois autres types erreurs de circonstance les plus fréquents. Ces derniers sont classés selon la fonction ou le type d'information qu'ils apportent à la phrase : (i) la manière {7,78%} au sens LSF (à savoir, le degré ; la comparaison, le moyen mis en œuvre, ou la qualité), (ii) la portée {5,91%} (en termes de distance [*répondant donc la question, à quelle distance ?]* ; la durée ; la fréquence) et enfin (iii) la location spatio-temporelle {4,67%} (à savoir répondant aux questions, où ? et quand ?). Imaginons donc un instant un texte d'une certaine linéarité dont le scripteur présente des informations, sans pouvoir bien préciser la portée de l'information, la manière ou le degré d'implication d'une action, voire sans pouvoir apporter d'explication sur la manière dont le lecteur pourrait "comprendre" les arguments ou informations présentés. Ce scénario est en quelque sorte le résultant de ces manquements observés au niveau 'circonstanciel'.

Quant aux signalements portant sur les procès et les participants, les résultats sont a priori sensiblement les mêmes que dans le premier volet, ce qui pourrait suggérer que le problème n'est pas simplement d'ordre grammatical et pourrait d'autant plus provenir d'une sorte de grammaire ou d'interlangue intermédiaire. Cependant, quand on examine de manière approfondie les erreurs annotées ici, on remarque que la plupart des erreurs de participants renvoient aux erreurs de références ou de choix sémantique tandis que celles annotées comme des erreurs de procès renvoient dans la quasi-totalité aux erreurs de progression temporelle – ce qui permet de changer notre regard sur ces signalements. Cela étant, il convient d'une part de souligner à ce niveau d'analyse que les signalements portant sur les procès et les participants ne sont pas particulièrement significatifs ou concluants. Tandis que, d'autre part, les signalements sur les différents types de circonstances demeurent très instructifs, puisqu'ils portent vraisemblablement sur les valeurs et fonctions intrinsèques de ces items qui ne sont pas maîtrisées par nos apprenants.

6.2.3 Le schéma interpersonnel appliqué aux erreurs d'acceptabilité textuelle

Les ré-annotations avec le schéma interpersonnel ont apporté leur lot d'éclaircissements : certains constats issus des sections précédentes ont été confirmés et précisés davantage tandis que d'autres points ont été mis en avant de manière inattendue. En effet, dans la figure 35 les résultats illustrés – au tout premier niveau du schéma d'annotation – peuvent sembler anodins. On a même l'impression qu'ils suivent en quelque sorte la distribution normale de la phrase simple : avec 29% d'erreurs signalées en « mood-1 », 50% en « residue » et 20% en dehors de ces deux catégories. On pourrait même être tenté d'affirmer que cette distribution est logique puisque le « residue » est la partie en dehors de la structure « mood+residue » où il y a le plus de « mots » ou de « place » pour

l'explication du message et donc de l'observation de phénomènes linguistiquement intéressants. Toutefois on est bien loin de ce scénario.

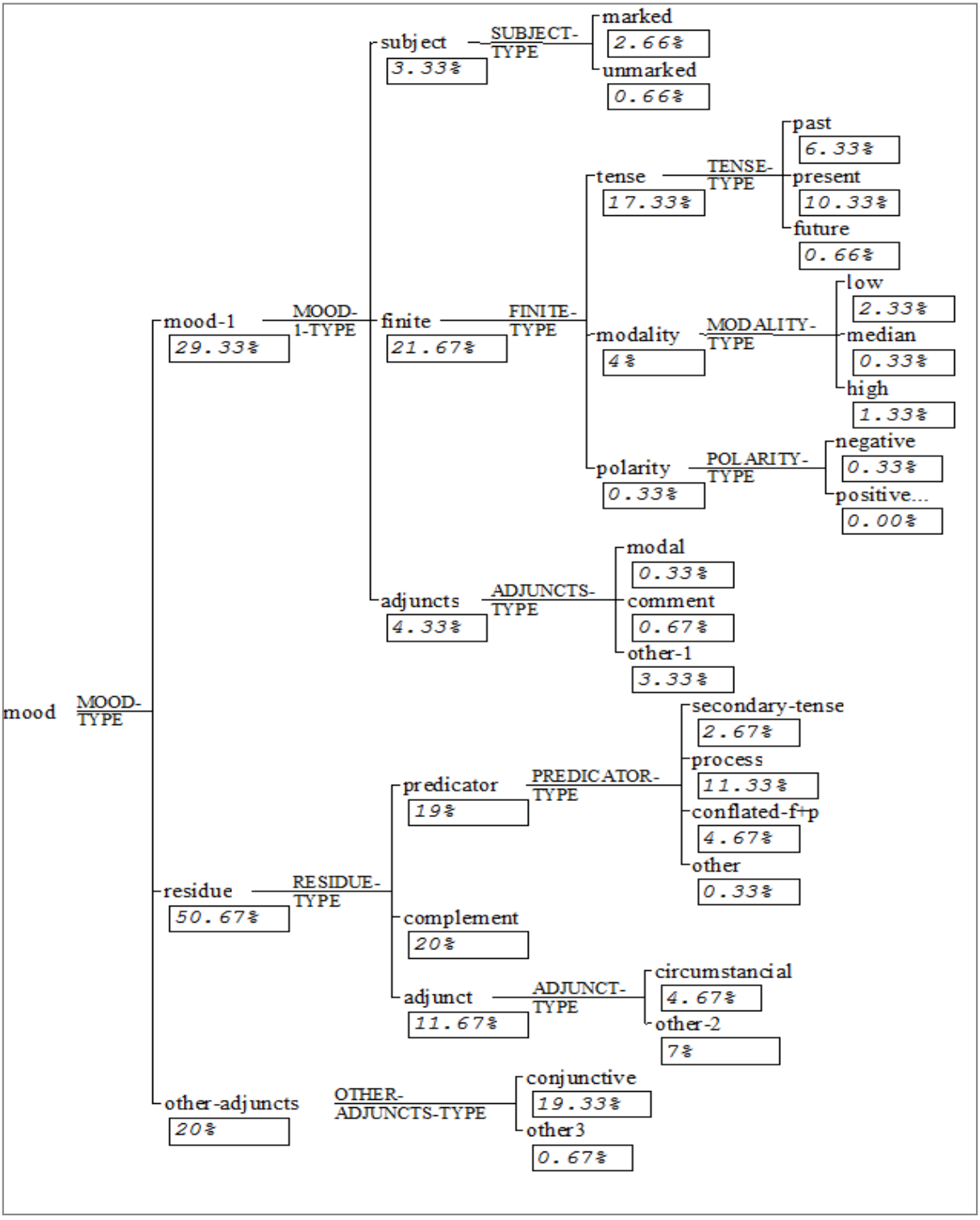


Figure 35 : Le schéma interpersonnel au service des erreurs textuelles

La répartition des erreurs à ce niveau se présente de la manière suivante : (a) 40% portant sur le « finite » et le « predicator » (respectivement 21% et 19%) ; (b) 35 % sur les ajouts tout type confondu (« mood-1 adjuncts » [4%], « residue adjuncts » [12%] et « other-aduncts » [20%]) ; (c)

20% sur le complément ; et (d) 3% sur les sujets. Examinons maintenant tout cela de plus près. Tout d'abord un cinquième des erreurs a donc été identifié comme ayant une fonction dite de conjugue[107] (*finite*, en anglais). Ce premier problème s'explique par l'emploi erroné du temps grammatical présent – pour la majorité des cas signalés – alors que le texte exige un temps grammatical différent. Ceci fait écho aux nombreuses erreurs de progression temporelle signalées en section 7.1. A cela s'ajoute l'emploi erroné des modaux auxiliaires en lieu et en place d'un autre, lui-aussi exigé par l'environnement textuel. Un exemple est fourni ci-après.

49. Last but not least the world of work has changed and is still changing, who can imagine today a post graduate who **could* can not speak [E]nglish or use a computer? (txt_091_sm2)

On voit donc très clairement dans la figure 35 le type de modalité[108] qui pose le plus de problèmes (à savoir dans un ordre décroissant « low : could » ; « high : will » et « median : would »). Mais tout cela est bien entendu à relativiser étant donné le faible effectif observé dans cette catégorie. Le deuxième problème identifié au niveau interpersonnel porte sur les erreurs de « prédicateur » qui constituent, elles aussi, un cinquième des erreurs comptabilisées. Hormis les erreurs de progression temporelle, ces dernières renvoient principalement à des d'erreurs d'ordre purement sémantique (souvent à la limite de la frontière des items collocationnels[109]).

Ce dernier point constitue un problème qui semble véritablement durable puisque ces erreurs d'ordre purement lexical et sémantique sont présentes dans chacune des classes grammaticales examinées dans le premier volet et elles sont également présentes parmi les catégories décrites en section 6.2.1. Cependant, il convient de souligner que, contrairement à la plupart des autres catégories existantes, les erreurs sémantiques ne peuvent être simplement corrigées par l'explicitation de règles grammaticales. Ces problèmes traduisent à notre avis, non pas un simple manque de vocabulaire, mais comme nous le verrons dans la section 8.2.2.1, une méconnaissance des rapports collocationnels privilégiés entre certains termes.

50. So women depend on stereotypes which are no realistic. Some changes have been **tried* $made$ but it's a deeper revolution that could change the situation. (txt_100_sm2)
51. Some countries **gather the reduction* $##$ to their emissions and the compensation: eco-tourism is one of them. (txt_117_sm1).

[107] Pour rappel le terme conjugue renvoie à « la partie 'conjuguée' du verbe […] qui attire les marques de temps et de nombre » et la négation (Banks 2005). Cf. section 3.2.5 pour une explication plus approfondie.
[108] Cf. la section 3.2.5 où les trois types de modalités ont été brièvement introduits ou voir Eggins (2004).
[109] Cf. la section 8.2.1 pour une discussion sur le rapport collocationnel ou phraséologique de certaines erreurs.

L'exemple (50) met en avant une sorte d'association collocationnelle impropre que l'on pourrait reformuler par « **to try some changes* ». Bien que syntaxiquement corrects, le verbe et son complément se heurtent à une certaine non-réciprocité collocationnelle. Sur le plan sémantique « try » renvoie à une notion de test ou à une phase expérimentale, tandis que « change » signifie (i) passer d'un état à un autre ou (ii) subir des modifications – ce qui suppose d'emblée un certain achèvement. Il existe donc des rapports unités proprement collocationnelles qui se rapprochent et témoignent de ces valeurs sémantiques : on dirait (i) « to make some change » ou (ii) « some changes have been made ». Au vu donc de l'exemple (50), sans oublier bien entendu les erreurs sémantiques de la section 6.2.1, nous pouvons soutenir que l'apprenant n'est pas vraisemblablement au courant ou ne s'est pas encore approprié ces valeurs collocationnelles – qui doivent s'ajouter à la simple connaissance des unités lexicales.

Le constat est indubitablement le même pour l'exemple (51), mais on remarque en plus que l'ensemble conduit à un non-sens général : d'où l'impossibilité de proposer une suggestion de modification. Précisons, par ailleurs, pour ce qui est l'apprentissage des rapports collocationnels qu'il incombe plus à l'apprenant d'acquérir progressivement des éléments lexicaux qu'à l'enseignant de lui fournir et expliciter les unités lexicales les unes après les autres. Et ce, étant donné le nombre important de collocations possibles par unité lexicale individuelle, il nous parait impensable d'imaginer que l'enseignant de langue étrangère soit obligé de s'acquitter de cette tâche colossale (en fournissant une liste encyclopédique de toutes les collocations connues). Notamment si ce dernier doit s'y prendre de la même manière qu'il enseigne par exemple les rapports syntaxiques privilégiés entre une proposition et un verbe, par exemple.

Le troisième problème relevé par le schéma interpersonnel porte sur les erreurs d'ajout. De plus, la nature de ces ajouts a été confirmée et plus encore précisée – comme illustré sur le schéma – à travers les 19% d'erreurs correspondant aux fonctions dites « conjonctives ». Ces résultats renvoient en grande partie aux erreurs de cadrage et aux erreurs de circonstances – comme nous l'avons démontré en section 6.2.1 et 6.2.2. Et le pourcentage a, en quelque sorte, conforté notre analyse précédente par le fait que Halliday & Matthiessen (2004) voient dans ces fonctions une relation de cohésion sémantique – qui se manifeste de manière prédominante dans la fonction dite de thème (et donc en début de phrase, cf. nos erreurs de cadrage). De ce fait, nous pouvons réaffirmer notre postulat selon lequel les apprenants auraient une réelle difficulté qui dépasse la simple maturité syntaxique et qui, de plus, touche tout particulièrement à l'assemblage des unités permettant de donner de la « texture » au texte.

Quant à notre quatrième et dernier problème, les points (c) et (d) sont examinés ensemble en raison notamment du fait qu'il s'agit d'items étiquetés en section 6.2.2 comme étant des participants (ou PR1 et PR2, comme nous les avons rebaptisés en section 5.2.4). L'éclaircissement apporté par le schéma interpersonnel permet de démontrer clairement que ce n'est pas tant la fonction « sujet » ou « PR1 » qui pose problème – du moins pas autant que la fonction « PR2 » qui regroupe plus de six fois le nombre d'erreurs identifiées en PR1. Mais au-delà de cette précision sur le nombre d'erreurs sur le PR2 par rapport au PR1, le schéma ne fournit pas davantage d'informations sur la nature exacte. Il est donc nécessaire de croiser ces résultats avec le schéma présenté pour comprendre que les erreurs de PR2 sont principalement sur les participants matériels (dits *scope* et *attribute*, en anglais) ou les participants relationnels (*carrier* et *token*).

De manière générale, le bilan du schéma interpersonnel est globalement très positif notamment par rapport aux éclaircissements apportés sur les différents types et fréquences d'erreurs observés dans l'emploi des temps grammaticaux et les différents modaux. A cela s'ajoute (i) la différenciation obtenue avec les erreurs observées sur le prédicateur ou le verbe lexical, par rapport au conjugue (*finite*) ou l'auxiliaire ; (ii) sans oublier, bien entendu, la confirmation des difficultés observées dans l'emploi des ajouts et des différentes fonctions circonstancielles rapportés précédemment dans la section 6.2.2.

6.2.4 Le schéma textuel appliqué aux erreurs d'acceptabilité textuelle

Parmi les trois schémas LSF utilisés pour effectuer les ré-annotations dans le deuxième volet de notre étude, le schéma textuel est celui qui a recueilli le moins d'informations à la fois qualitatives et quantitatives. Mais malgré le fait que ces résultats étaient d'une certaine manière attendue, nous pensions que ce schéma (cf. figure 36) se serait révélé plus pertinent ou du moins plus à même d'indiquer en son sein un plus grand nombre d'erreurs d'acceptabilité textuelle. Toutefois cela ne fut pas le cas. Nous nous sommes rendu compte, par conséquent, que la dimension que nous croyions proprement textuelle n'était pas suffisamment élargie, ne serait-ce que par la portée restrictive que nous avons adoptée en analysant uniquement la fonction « thème » au détriment du « rhème ». Néanmoins nous avons tout de même réussi à en tirer profit, en relevant les trois constats suivants.

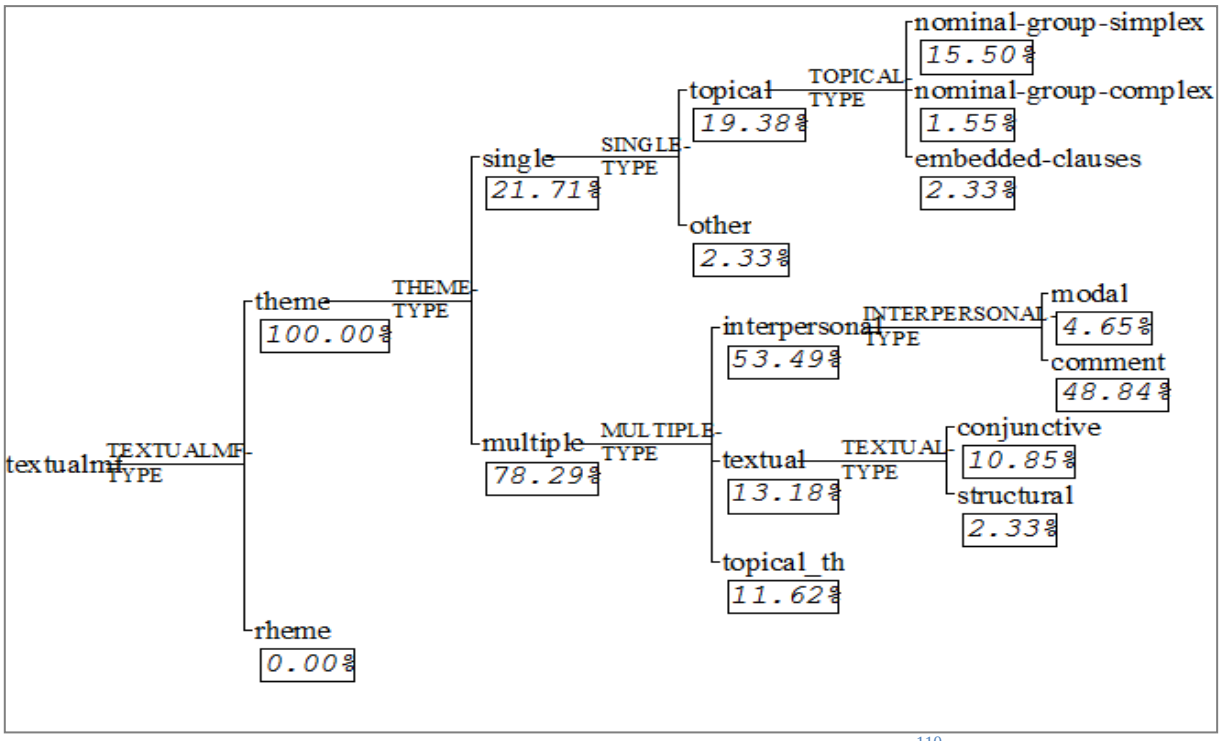


Figure 36 : Le schéma textuel au service des erreurs textuelles[110]

a) Le « single-theme » constitue la seule véritable observation significative de ce schéma puisqu'il montre la fréquence réelle des erreurs portants sur la fonction de « thème-topical » (ou sujet grammatical, en grammaire traditionnelle) en précisent que ce sont les groupes nominaux simples qui enregistrent le plus d'erreurs et non les groupes nominaux complexes. On aurait pu s'attendre à un résultat inverse, notamment si l'on estime que les groupes complexes demandent une maturité syntaxique plus que considérable de la part de nos apprenants. Mais nous pensons que ces chiffres n'indiquent que la fréquence globale d'emploi – ce qui signifie que les apprenants auraient développé des stratégies d'évitement et privilégieraient la construction des groupes syntaxiques simples à la place des groupes complexes.

b) Le deuxième constat n'est pas nouveau en soi mais vient tout simplement conforter notre « diagnostic » sur la difficulté qui, pour les apprenants, consiste à réaliser la cohésion interne de leur texte. En effet, si l'on se réfère de nouveau à la figure 36 la création des « thèmes-multiples » est particulièrement problématique : notamment les thèmes dits interpersonnels qui renvoient aux moyens employés pour faire émerger « la voix du scripteur » avec son angle ou perspective argumentative (du moins, pour ce qui est des textes dans notre corpus). Et c'est justement ici que l'on observe le plus d'erreurs – comme nous l'avons démontré dans

[110] Toutes les catégories n'ayant pas enregistré un pourcentage supérieur à 0 ont été écartées du schéma, à l'exception, bien entendu, du rhème.

les deux sections précédentes. Pour ce qui est de l'appellation « thème-textuel », nous rappelons que « *conjunctive* » renvoie majoritairement aux ajouts conjonctifs (cf. section 3.2.5) tandis que « *structural* » renvoie aux conjonctions de coordination et de subordination. Selon Halliday & Matthiessen (2004), les premiers ont une fonction de cohésion sémantique et les derniers une fonction structurelle. Mais en parcourant les éléments en contexte réel, il nous paraît arbitraire de vouloir séparer de manière définitive ces deux éléments en raison du fait qu'ils réalisent tous les deux une fonction réelle de manière simultanée (i) dans la construction de l'objet texte et (ii) dans l'émergence de l'angle d'approche du scripteur.

c) Enfin, *a posteriori*, nous pensons que les erreurs du rhème auraient dû être signalées au même titre que celles observées dans le thème. En effet, considérons le constat mis en avant dans la figure 35 : le schéma interpersonnel a permis de démontrer que (A) 70,67% des erreurs comptabilisées ont été signalées dans (i) le résidu et (ii) les ajouts en position post-verbale (ou plus précisément à l'extrême périphérie droite de la phrase) et (B) uniquement 29,33% des erreurs ont été identifiées sur le mode et la conjugue. Si nous croisons donc ces données avec le schéma 36 ci-dessus, l'ensemble des erreurs identifiées en position de thème est inférieur à 29,33%. Cela signifie que le niveau textuel, tel que nous l'avons analysé, s'avère très parcellaire. Nous aurions gagné par conséquent à approfondir la structure rhématique (même si la théorie de la linguistique systémique fonctionnelle, elle-même, n'a pas encore établi de composants à l'intérieur du rhème). L'avantage aurait été d'avoir une analyse aussi fine que celle obtenue dans le thème.

6.3 Le bilan des erreurs textuelles

En définitive, le deuxième volet d'annotation de notre étude s'articule autour des erreurs repérées et étiquetées à un niveau proprement textuel : à savoir, là où la grammaticalité de la phrase n'a pas été mise en cause – mais où l'environnement textuel immédiat a rendu les items sélectionnés inacceptables. Rappelons ensuite que les sept catégories d'erreurs d'acceptabilité textuelle ne sont pas exhaustives et renvoient tout simplement aux traits distinctifs les plus fréquemment observés dans ce deuxième volet d'annotation. De plus, comme nous l'avons soutenu en section 6.2.2, les erreurs relevées ici font rarement l'objet d'une réflexion didactique propre et ne sont, de ce fait, pas corrigées de façon ciblée en cours de langue. Nous pensons que ce manque d'attention particulière risque de retarder l'aptitude des apprenants à s'auto-corriger voire à provoquer un effet de

fossilisation si ces derniers n'arrivent pas à identifier ce problème et à y remédier au moment opportun.

Les ré-annotations des items obtenus en section 6.2.2 avec les trois schémas issus de la linguistique systémique fonctionnelle appellent plusieurs remarques. Tout d'abord chacun des schémas a permis de mettre en lumière sur des phénomènes linguistiques spécifiques – nous permettant de la sorte de confirmer davantage les étiquetages de la section 6.2.2 (notamment en indiquant les spécificités des différentes erreurs à trois niveaux d'analyse différents). Il convient de signaler cependant qu'aucun des trois schémas n'a été capable de ré-étiqueter l'ensemble des 684 annotations obtenues en section 6.2.2 : à titre d'exemple, seul 47% a pu être ré-annoté avec le schéma expérientiel ; 46% avec le schéma interpersonnel ; et enfin uniquement 19% avec le schéma dit textuel. Ce constat appelle deux remarques supplémentaires : malgré le fait que chacun des schémas a pu nous apporter des éclaircissements sur les caractéristiques distinctives des erreurs, (i) les résultats des trois schémas LSF sont tous à tempérer au vu du faible effectif ré-étiqueté et (ii) ces derniers ne constituent pas en soi des modèles capables de rendre compte de tous les phénomènes initialement observés à ce niveau d'annotation.

Hormis ces rappels d'ordre général, il convient de souligner également ici que nous avons rencontré – aussi bien dans les erreurs textuelles du volet 1 (cf. section 6.1) et les erreurs d'acceptabilité textuelle du volet 2 (cf. section 6.2) – un point particulièrement délicat. Ce point constitue une sorte de limite méthodologique et conceptuelle En effet, la caractérisation des erreurs comme étant proprement textuelles ou relevant du système linguistique a été une entreprise complexe, en raison notamment du fait qu'un nombre important des occurrences erronées identifiées dans ce chapitre semblait appartenir conjointement aux deux types d'erreurs cités ci-dessus. Et malgré le fait que nous avons essayé de ne retenir que des énoncés grammaticalement corrects pour le volet 2 (cf. section 6.2), des chevauchements ont tout de même été observés – notamment dans les erreurs référentielles, les erreurs de cadrage et une des sous-catégories d'erreurs de focus dite de progression temporelle (ceci concerne ici aussi bien les auxiliaires que les modaux).

Nous pensons donc que ces erreurs – qui renvoient dans une certaine mesure à la structuration du texte et tout singulièrement à la structuration et à la continuité informationnelle – revêtent un certain chevauchement inhérent de par leur nature et leur fonction textuelle. Mais nous concédons que des études supplémentaires sont nécessaires pour comprendre davantage pourquoi nous observons un certain nombre de chevauchements sur des erreurs précises et pour mieux étudier ces

chevauchements afin de faire émerger un moyen fiable de délimiter les erreurs du système linguistique par rapport aux erreurs d'acceptabilité textuelle.

(Chapitre VII) L'influence du temps et des contacts linguistiques sur les erreurs

Ce chapitre, à l'inverse des deux précédents, a pour but d'ouvrir la discussion sur nos principaux résultats. Tout d'abord, en section 7.1, nos trois profils linguistiques ou « groupes tests » (E-CLR, F-CLR et F+CLR) (cf. Annexe A2) sont examinés à la lumière de deux variables : l'évolution des occurrences erronées entre le premier (sm1) et deuxième (sm2) semestres et la durée des différents contacts linguistiques propres à chaque profil. L'incidence entre ces deux variables est explicitée à travers une synthèse croisée avec les résultats obtenus dans les deux volets d'annotation (à savoir, les chapitres V et VI respectivement). Nous verrons par ailleurs que l'ensemble des résultats conforte notre postulat du départ sur les erreurs à la fois du systémique linguistique et celles que nous appelons acceptabilité textuelle : plus précisément que ces deux types d'erreurs ne sont pas régis par les mêmes maîtrises, compétences et connaissances et n'évoluent pas, par conséquent, de la même manière au fil du temps.

Dans la section 7.2, l'impact de la langue maternelle est passé en revue. Il est tout particulièrement question de délimiter une frontière entre les erreurs qui peuvent être attribuées à la langue maternelle, ou à l'interlangue en construction ou celles qui ne peuvent être explicitées ni par le premier cas de figure ni par le deuxième. Une attention est également portée sur les différents types de transfert que l'on peut identifier sur des phénomènes quasi-identiques. Nous verrons par la suite que le fait d'identifier des occurrences comme étant la manifestation d'interférence ou de transfert et d'établir un lien de causalité distinctif entre les différentes influences de la langue maternelle relève d'un procédé plus complexe qu'il n'y paraît. Nous appelons enfin à une plus grande attention dans la caractérisation de la notion de transfert afin d'éviter que ce terme ne devienne un fourre-tout.

7.1 L'incidence des contacts linguistiques

De manière générale, il est admis que la motivation personnelle aussi bien que le milieu social ou le contexte institutionnel de l'apprentissage sont des facteurs importants à prendre en compte dans l'acquisition d'une langue étrangère (cf. Dörnyei 1998 ; Ushida, 2005 ; Taguchi, 2008). Mais étant donné que le premier relève d'une notion subjective[111] et les deux derniers n'ont pas fait l'objet d'une attention particulière dans la sélection des sujets-participants de notre étude, il a fallu identifier un autre moyen d'étudier à la fois de manière concrète et objective les différences qui peuvent influer sur l'écart de performance observé dans la production écrite de nos apprenants. Nous avons, de ce fait, choisi d'étudier la durée et les types de contacts linguistiques observés chez ces derniers avec pour objectif principal de voir si ces deux éléments ont véritablement influé (et si oui, à quel degré ?) sur les résultats obtenus dans les chapitres V et VI.

Cela étant dit, rappelons que la durée et les différents contacts linguistiques observés chez les sujet-participants ne sont pas homogènes, comme nous l'avons démontré en section A2.1.2 de l'annexe. Nous notons toutefois que certains traits distinctifs sont communs à la trajectoire linguistique de beaucoup de nos apprenants, par exemple le fait que trois quarts ont commencé l'apprentissage de l'anglais en même temps – à savoir à l'école primaire. Les résultats obtenus des croisements de regard sur ces éléments sont, de ce fait, discutés dans les deux sous-sections qui suivent. En section 7.1.1 nous allons nous intéresser à ceux d'entre eux qui se trouvent aux deux extrémités de la variable temporelle, à savoir ceux qui ont eu une initiation précoce à la langue anglaise comparés à ceux qui ont eu une initiation plus tardive, c'est-à-dire, respectivement, ceux qui ont commencé à la maternelle (17 sujets-participants) et au collège (16 sujets-participants). Ensuite en section 7.1.2, nous nous interrogeons sur l'apport des séjours prolongés en pays anglophones et leur impact potentiel sur les trois groupes tests, tels qu'ils sont établis en section A2.2 de l'annexe.

7.1.1 Commencer l'anglais à la maternelle ou au collège : quels avantages ?

Comme nous l'avons précédemment signalé, la majorité de nos sujets-participants ont commencé l'apprentissage de l'anglais au primaire ; nous avons cherché, par conséquent, à savoir si ceux qui auraient commencé à une période différente auraient des résultats distincts que l'on pourrait

[111] Le fait de vouloir réussir une matière ne signifie pas que l'on y arrivera ; ce constat est également valable pour l'apprentissage des langues étrangères. Donc, au vu des travaux cités et malgré le fait d'accorder une importance non-négligeable à la motivation, nous l'avons exclu comme facteur à prendre en compte. Cela étant, nous soulignons également le fait que mesurer la motivation recouvre une certaine subjectivité, et étant donné que nous ne disposons pas des outils nécessaires pour mener à bien et de façon complète une telle analyse, nous avons jugé judicieux d'écarter cette variable de l'étude.

identifier de manière statistique. Cela étant, nous avons effectué un *test de t* avec ceux qui ont commencé soit avant soit après l'école élémentaire et le résultat est sans appel : avec une valeur *p* de 0,9956[112], nous pouvons affirmer qu'il n'y a pas d'écart statistiquement significatif entre les résultats obtenus pour ces deux « sous-groupes tests », ni dans l'annotation des erreurs dites du système linguistique ni celles d'acceptabilité textuelle, signalées respectivement ER.ling (volet 1) et ER.text (volet 2) dans le graphique ci-après. Autrement dit, ce premier constat suggère à première vue qu'il n'y aurait pas de différence réelle et significative entre le fait de commencer l'apprentissage de l'anglais à la maternelle et au collège.

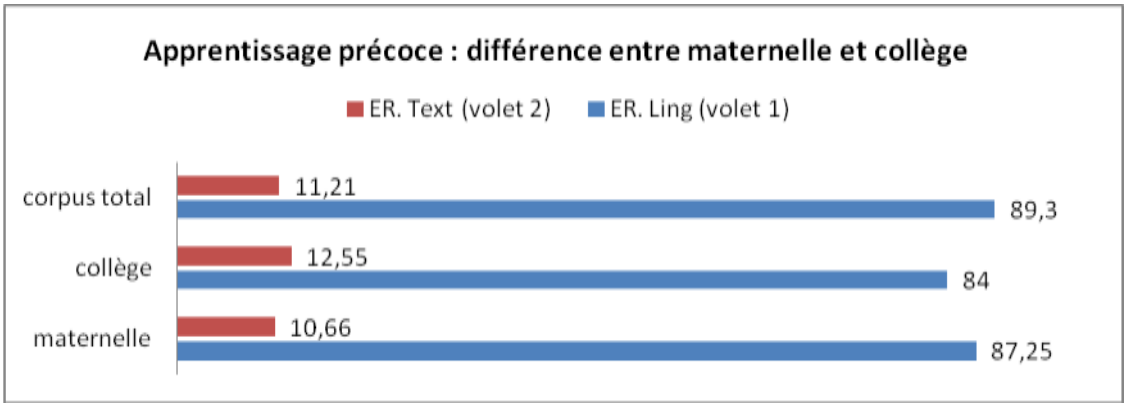


Figure 37 : l'influence d'un apprentissage précoce

La figure 37 met en avant en effet la moyenne des erreurs observées selon l'établissement dans lequel les apprenants auraient commencé l'apprentissage de l'anglais. La moyenne des 122 participants est également présentée dans « corpus total ». Pour écarter tout biais dans les résultats, plusieurs manipulations ont été effectuées par la suite dans le but d'identifier toute différence possible au niveau « intra-groupe ». Par exemple, tous les sujets-participants ayant été scolarisés uniquement en France et ayant commencé l'anglais à la maternelle ou au collège ont été comparés entre eux pour ensuite les comparer à ceux ayant signalé avoir vécu ou avoir été scolarisés à l'étranger : il est en ressorti une différence minime ; p = 0,9956 pour le groupe ensemble ; p = 0,9901 pour ceux ayant un parcours linéaire en France et p = 0,9783 ceux ayant eu un parcours hors hexagone. L'ensemble des écarts observés ne recouvre, par conséquent, aucune valeur significative.

Ce résultat est inattendu et ne corrobore pas les hypothèses de la période critique qui, rappelons-le, sous-entend que plus l'apprentissage est précoce, plus l'effet sera visible sur le long terme. Il convient également de préciser ici que l'écart en termes d'années entre l'école maternelle

[112] Pour rappel, quand la valeur *p* est égale ou supérieure à 0,5 (≥) il est admis que l'écart observé n'est pas significatif tandis que si elle est inférieure à 0,5 (≤) l'écart est considéré comme significatif.

(l'inscription possible dès 3 ans jusqu'à 5 ans) et le collège (où l'inscription moyenne est à partir de 11 ans) n'est pas négligeable. Cela étant dit, nous sommes néanmoins conscient que plusieurs facteurs méritent d'être soulignés avant toute généralisation : nous en présentons deux. Premièrement, il peut y avoir une corrélation entre la longueur d'un texte et le nombre d'erreurs comptabilisées. Par exemple les étudiants intermédiaires (B2) ou avancés peuvent être amenés à écrire deux ou trois fois plus que les étudiants « débutants » (A1, A2). Donc la probabilité d'avoir des erreurs augmente avec la longueur du texte. A cette première remarque s'ajoute le fait qu'un texte jugé excellent en classe de langue étrangère ne signifie pas qu'il est dépourvu d'erreurs, de même que le nombre d'erreurs signalées dans un texte donné ne peut être le seul moyen de juger de sa qualité en tant qu'un ensemble. Il est donc important de concéder que l'écart entre les deux sous-groupes peut se manifester différemment dans leurs productions écrites.

7.1.2 Séjourner en pays anglophones : quel bilan ?

Après avoir examiné la première variable temporelle de notre étude en termes de contact précoce, passons maintenant à la deuxième variable temporelle qui porte sur le contact tardif et plus particulièrement la durée de ce contact et son influence sur les trois profils linguistiques différents de nos apprenants (cf. section A2.2 de l'annexe). Pour rappel, les profils sont reprécisés ci-dessous.

- [P1] ou (E+1CLR) fait allusion aux participants étrangers (tout type confondu) ayant suivi la plus grande partie de leur scolarité à l'étranger.
- [P2] ou (F-CLR) désigne les participants francophones – qui sont nés ou ont passé toute leur vie en France – et qui n'ont ni suivi des cours entièrement en anglais ni séjourné dans un pays anglophone. Les participants retenus dans cette catégorie ne parlent que le français à la maison.
- [P3] ou (F+CLR) renvoie à ceux qui ont un contact linguistique régulier en dehors du cadre institutionnel et qui ont, de plus, séjourné dans des pays anglophones dont la durée totale est supérieure à deux mois.

Rappelons également que le raisonnement, à partir de ces trois profils, est de mesurer dans un premier temps l'impact des différents contacts linguistiques sur les erreurs relevées dans notre corpus et dans un deuxième temps d'établir des corrélations possibles entre les fréquences statistiques observées et les différences typologiques propres à chaque groupe. En termes de méthodologie, l'ensemble des erreurs signalées dans les chapitres V et VI ont été regroupées selon le profil linguistique du sujet-participant qui les aurait commises. L'évolution des erreurs entre les deux semestres d'étude a donc été établie pour chaque participant. Le tableau 37 suivant donne un

exemple pour des participants E+CLR. A titre d'illustration, le tableau indique que *sp19* appartient au profil n°1 [P1] ou (E+1CLR) et il est passé de 54 erreurs en sm1 à 40 en sm2, avec une réduction de 25%.

	sp	sm1	sm2	(%)	
E+CLR	sp19	54	40	-25,93	↘
E+CLR	sp28	79	69	-12,66	↘
E+CLR	sp33	21	18	-14,29	↘
E+CLR	sp52	49	30	-38,78	↘
E+CLR	sp61	26	13	-50	↘

Tableau 37 : Evolution des participants (E+CLR)

Passons maintenant aux résultats qui nous intéressent. Comme nous l'avons signalé dans le chapitre VI, il y a eu une diminution globale de 13,8% entre les erreurs du système linguistique (volet 1) relevées en sm1 par rapport au sm2. Force est de constater cependant que la distribution de cette réduction n'est pas homogène, et tout singulièrement ne se manifeste pas de la manière escomptée pour les trois profils établis. La tendance est illustrée dans la figure 38 ci-dessous. Tous les sujets-participants (sp) E+CLR (sp19 à sp61) et F-CLR (sp01 à sp31) ont vu leur nombre d'erreurs réduites en sm2 par rapport au premier semestre de -12% jusqu'a -50% pour le premier groupe et -20% jusqu'à -63% pour le deuxième. Leur moyenne respective est de -28% et -43%. Ces résultats étaient dans une large mesure attendus tels quels puisque que, de manière intuitive, on s'attend à ce que les apprenants, tout niveau confondu, fassent moins de fautes au fur et à mesure qu'ils progressent en cours de langue.

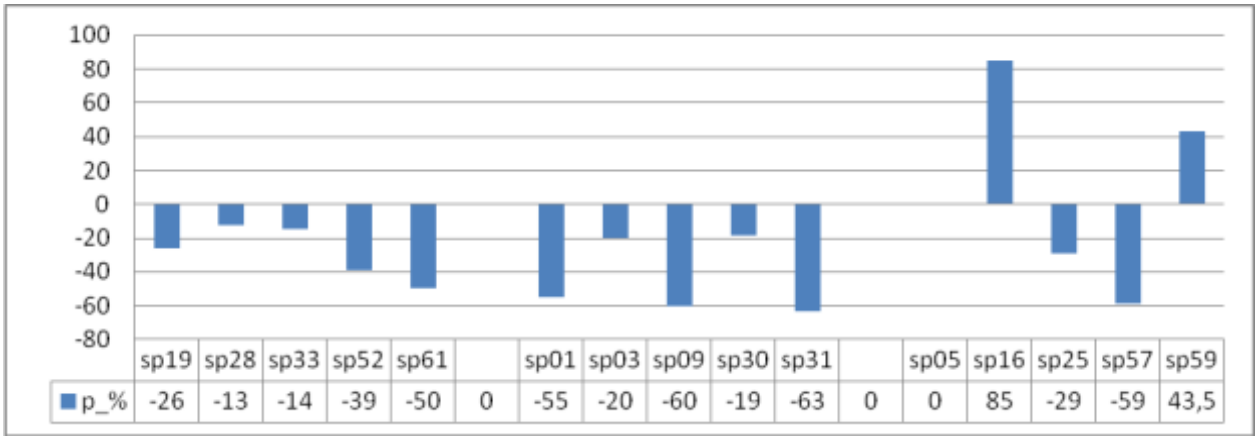


Figure 38 : L'évolution semestrielle (E+CLR, F-CLR, F+CLR) (volet 1)

Le constat n'est cependant pas le même pour les cinq sujets-participants à l'extrême droite de la figure 8.1.2-1 (sp05 à sp59) qui correspondent au profil F+CLR. Rappelons à nouveau que ces apprenants ont eu un contact linguistique jugé important avec l'anglais en dehors du cadre proprement institutionnel – en considérant bien entendu leur deux mois minimum de séjours

effectués dans des pays anglophones. La tendance observée dans ce groupe va d'une réduction de moins -58% à une augmentation de +85%, avec une moyenne donc de +8%. Cette contre-tendance par rapport aux deux groupes précédents est pour le moins totalement inattendue, d'autant plus que les deux participants ayant enregistré les plus fortes augmentations d'erreurs ont effectué des séjours plus longs que la moyenne de ce groupe (à savoir plus de 4 mois pour sp16 et 1 an pour sp59.

Cette contre-tendance inédite semble traduire à notre sens non pas une détérioration des compétences acquises mais au contraire l'augmentation des stratégies personnelles de rédaction qui visent à l'amélioration de la maturité textuelle, tout en augmentant « les prises de risque ». En effet, nous pensons que les apprenants dont la maturité syntaxique et textuelle n'est pas encore attestée adopteront de manière générale des stratégies d'évitement (cf. Jin 2000 et Vasquez 2005). C'est-à-dire que ces derniers se limiteront à des structures déjà acquises et maîtrisées en langue maternelle ou à des structures nouvellement apprises en cours de langue étrangère. Tandis que ceux ayant un contact linguistique fréquent (F+CLR) - et donc un *output* et un *input* en langue étrangère qu'ils jugent davantage solides – chercheront à mettre en place des nouvelles stratégies et à « s'impliquer » davantage dans leur rédaction. L'implémentation de ces stratégies se traduira, de ce fait, par des textes plus longs (ceci renvoie au constat fait en section 7.1.1) : certaines stratégies seront tout à fait adaptées au système linguistique ou au genre textuel/contexte de rédaction, tandis que d'autres ne le seront pas. A titre d'exemple, ce qui conforte davantage cette hypothèse est l'augmentation chiffrée observée dans les erreurs dites grammaticales mais également celles dites de mise en phrase à un niveau intra-groupe chez les participants F+CLR.

Mais qu'en est-il des occurrences que nous avons appelées des erreurs d'acceptabilité textuelle ? Tout d'abord, la tendance générale observée à ce niveau n'est pas à la baisse comme c'est le cas avec les erreurs du système linguistique mais augmente de manière globale de 25% – pour ce qui est de nos trois groupes tests. Cependant l'augmentation constatée est loin de constituer une distribution régulière entre les différents participants et tend plus à la fluctuation généralisée, comme illustrée dans la figure 39 ci-dessous. Il y a de nombreux écarts à la fois à un niveau intergroupe et intra-groupe. A titre d'exemple, dans le groupe E+CLR la tendance varie d'une diminution de -55% jusqu'à une augmentation de +100% - avec une moyenne de -9% ; dans le groupe F-CLR elle est de -57% jusqu'à une augmentation de +233% - avec une moyenne de +85% ; dans le groupe F+CLR la tendance oscille entre -75% et +150.

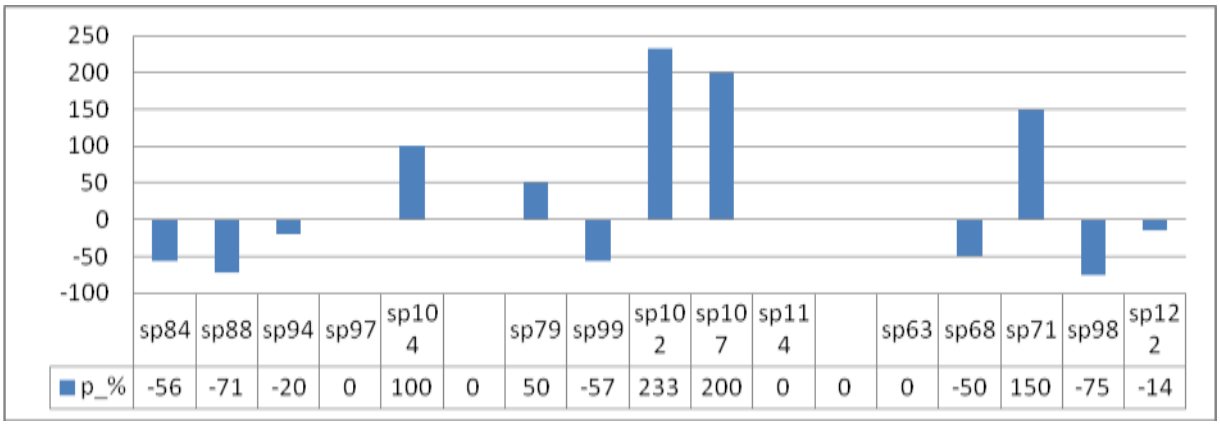


Figure 39 : L'évolution semestrielle des erreurs d'acceptabilité textuelle (E+CLR, F-CLR, F+CLR) (volet 2)

Ces premiers éléments indiquent que contrairement à la baisse généralisée de 80% observée dans les erreurs du système linguistique (volet 1, cf. section 7.1.1), la diminution réelle des erreurs d'acceptabilité textuelle n'est que de -47% tandis que l'augmentation est de +33% avec une fréquence inchangée chez 20% de l'ensemble des sujets-participants dans les trois groupes tests. À notre sens, cette fluctuation signifie d'une part que les deux types d'erreurs n'opèrent pas au même niveau dans l'ordre acquisitionnel de l'anglais langue étrangère, et d'autre part que le simple fait d'exposer des apprenants à des cours intensifs d'anglais de spécialité (comme ce fut le cas de nos apprenants) – sans explicitation ou métadiscours sur les éléments attendus lors d'une rédaction – ne suffisent pas pour améliorer ou « remédier » à l'ensemble des écueils proprement textuels relevés dans les productions écrites en langue étrangère. Nous pouvons affirmer cependant que les cours de spécialité suffisent, au vu des résultats dans le volet 1, pour produire une diminution rapide du nombre d'erreurs dites du système linguistique et qu'ils contribuent, de ce fait, indirectement à l'amélioration de la compétence langagière – en termes de maturité syntaxique et lexicale.

Examinons maintenant de manière plus approfondie les renseignements apportés par la figure 39. A titre d'information, le graphique renvoie à 15 sujets-participants qui sont regroupés par groupe de cinq. Le groupe de participants étrangers (E+CLR) correspond à sp84 jusqu'à sp104, le groupe d'apprenants sans contact linguistique avec l'anglais en dehors du cadre institutionnel (F-CLR) correspond à sp79 à sp114 ; et le groupe avec un contact linguistique renforcé (F+CLR) de sp63 à sp122. Si l'on regarde en dessous de l'axe horizontal, la diminution est présente dans chacun des trois sous-groupes mais demeure relativement stable voire minime face aux écarts importants observés dans les fluctuations au-dessus du même axe. Face à cette variable temporelle (c'est-à-dire, la durée des différents contacts linguistiques), ce qu'il convient de noter est que deux groupes qui ont fait l'objet de contacts linguistiques fréquents semblent « se comporter » de manière quasi-identique : à savoir le groupe de participants ayant été scolarisé à l'étranger et les participants

français ayant passé plus de deux mois dans des pays anglophones. En effet, à l'intérieur de chacun de ces deux groupes 60% des sujet-participants maintient une nette réduction dans leurs erreurs textuelles en sm2 par rapport au premier semestre ; 20% n'ont pas montré de changement statistique chiffré ; et les derniers 20% ont vu leur nombre d'erreurs augmenter au deuxième semestre.

Ces éléments de convergence laissent penser dans un premier temps que ces contacts renforcés entre langues ont un impact sur la réduction des erreurs textuelles en langue étrangère. Pour contraster ces éléments de rapprochement entre E+CLR et F+CLR, on pourrait signaler le fait qu'à l'intérieur du groupe F-CLR, la diminution observée n'est que de 20% : ce qui traduit une véritable différence entre nos trois groupes tests. Ce que l'on doit tout d'abord retenir de ces comparaisons est que face à un enseignement en anglais de spécialité, une très grande majorité des participants semble en avoir tiré profit en perfectionnant leur compétence langagière de base de manière rapide. Par contre, si l'on prend en compte le fait que les trois groupes tests ont participé au même cours d'anglais de spécialité pendant une année et que ceux qui sont sans contact linguistique régulier avec l'anglais en dehors du cadre institutionnel (F-CLR) ont enregistré de façon unanime la plus faible baisse d'erreurs textuelles, il convient de dire que l'impact du cours sur la compétence textuelle doit être tempéré et demeure par conséquent mitigé. La question que l'on se pose maintenant est de savoir pourquoi les étudiants ayant vécu à l'étranger et ayant eu un contact très important souvent avec deux ou trois langues (cf. section A2.1.3 de l'annexe) et les apprenants francophones ayant séjourné à plusieurs reprises ou de manière prolongée dans des pays anglophones arrivent à la même fréquence de réduction dans les erreurs textuelles. Cette question demeure pour l'instant sans réponse.

Pour tenter de trouver une réponse à cette question, poursuivons l'analyse en portant une attention particulière à nos neuf types d'erreurs d'acceptabilité textuelle – couplés avec la fréquence d'occurrence correspondante telle qu'elle a été relevée dans nos trois groupes tests. Ainsi, en commençant par le pourcentage total obtenu dans le tableau 38, notons une distribution qui parait plus au moins régulière vis-à-vis des trois groupes tests. Mais afin de ne pas tomber dans des généralisations fortuites et hâtives, il est important d'apporter deux éléments de précision : (i) soulignons les fortes disparités à la fois à un niveau intra- et intergroupe (c'est-à-dire selon les différents types d'erreurs et selon le groupe donné) et (ii) précisons également que l'ensemble des éléments dans le tableau ci-dessous sont à tempérer, notamment en raison du faible effectif retenu à

ce niveau (cf. chapitre VI, 684 occurrences) sans oublier bien entendu le fait qu'il y a des écarts réels entre les trois groupes tests et l'ensemble des autres sujets-participants de l'étude.

Erreur d'acceptabilité textuelle	E+CLR	F-CLR	F+CLR
Erreur référentielle	55,56	33,33	11,11
Erreur sémantique	39,39	24,24	36,36
Erreur de topicalisation	12,5	37,5	50
Erreur de coordination	22,22	33,33	44,44
Erreur de positionnement (ordre)	27,27	27,27	45,45
Erreur de mise en phrase	46,15	35,9	17,95
Erreur textuelle (ostension)	0	44,44	55,56
Erreur de progression thématique	0	100	0
Erreur de concordance des temps	47,06	23,53	29,41
total	37,67	32,19	30,14

Tableau 38 : La distribution des erreurs textuelles selon le profil linguistique (volet 2)

Cela étant, quelques brèves remarques s'imposent. Au vu du tableau 38 les erreurs de référence semblent constituer un problème majeur – de manière plus accentuée – chez les participants étrangers que chez les participants des deux autres sous-groupes. Ceci est également le cas pour les erreurs de concordance des temps ou encore les erreurs de mise en phrase, où ils représentent respectivement 47% et 46 % des occurrences signalées. Certaines erreurs semblent néanmoins le propre des apprenants francophones : (i) comme par exemple les erreurs d'ostension (c'est-à-dire, une erreur identifiée dans le fait de signaler un exemple ou apporter une clarification) ; (ii) de topicalisation ; (iii) et dans une moindre mesure les erreurs de coordination. Il en va de même pour les erreurs de progression thématique qui ont été signalées uniquement chez les participants ayant un parcours linéaire (F-CLR), mais nous nous gardons de tout commentaire sur ce résultat étant donné le très faible effectif observé dans cette catégorie, tout singulièrement dans les trois groupes tests.[113]

Il convient également de souligner le fait que certaines erreurs ont été annotées principalement chez les sujets participants ayant bénéficié d'un séjour en pays anglophones. Ces erreurs appellent deux remarques supplémentaires. Quand le groupe F+CLR se trouve en tête (en termes chiffrés) d'un type d'erreur spécifique, le groupe F-CLR se trouve trois fois sur quatre en deuxième position. Ce constat peut laisser croire dans un premier temps qu'il y a une caractéristique partagée entre ces deux groupes qui occasionnent ces erreurs spécifiques. La deuxième observation permet cependant d'écarter cette hypothèse, du moins pour l'instant. En effet, les cas d'erreurs dans lesquelles

[113] Précision à titre d'information que plusieurs erreurs de progression thématique ont été écartées de l'étude afin d'éviter ce que nous appelons une analyse de « double peine ». cf. l'introduction, chapitre VI.

F+CLR est majoritaire renvoient tous à l'idée de construction ou développement du texte, dans le sens d'un élargissement ou explicitation d'une information – notamment avec les erreurs de topicalisation et d'ostension. Ce constat suggère que le groupe F+CLR est davantage dans une dynamique de construction textuelle, par rapport aux participants F-CLR et qu'il est donc logique que ce groupe rencontre ces problèmes spécifiques. Mais le fait que les F-CLR se trouvent en deuxième position ne coïncide pas avec l'hypothèse que nous avons émise antérieurement en soulignant que ce groupe précis est vraisemblablement dans une stratégie d'évitement des situations complexes – en termes notamment d'emploi de syntaxe complexe ou le fait de tester de nouvelles stratégies dans les productions écrites en anglais. Ce « revirement » ne pourrait être complètement expliqué qu'à la lumière de nouvelles données davantage axées sur ces deux groupes précis.

Erreur d'acceptabilité textuelle (évolution globale)	
Erreur de topicalisation	+166,7%
Erreur de positionnement (ordre)	+120%
Erreur de concordance des temps	+112,5%

Tableau 39 : Les erreurs textuelles les plus « tenaces »

Quant aux erreurs textuelles qui émergent de nos analyses contrastives comme ayant augmenté et qui ne suivent pas, de ce fait, la tendance générale, elles sont au nombre de trois, à savoir les erreurs de topicalisation, de positionnement et de concordances de temps. Dans l'ensemble, ces trois erreurs traduisent des fonctions proprement textuelles, à savoir qu'elles contribuent à la structuration et la continuité informationnelle. La première des trois permet d'introduire et ensuite de « cadrer » les informations qui vont suivre dans un paragraphe donné ; la deuxième accentue cette notion de cadrer les informations présentées dans la mesure où elle porte sur la tentative du scripteur d'ajouter une précision singulière à un élément et c'est justement ces éléments de précision que l'apprenant n'arrive pas à bien « cadrer » ou à positionner. Pour ce qui est de la troisième erreur, elle renvoie à une attention particulière qu'il faut porter sur la transition entre les informations présentées. Elle constitue de ce point de vue un élément clé dans la structuration de l'information, de par sa fonction de « tisser » les énoncés entre eux aussi bien au niveau grammatical qu'au niveau sémantique. Cela étant dit, un texte dans lequel ces trois types d'erreurs sont abondamment présents, ce qui au vu des résultats pourrait être le cas pour un certain nombre de participants, traduit un écrit, à notre sens, dont le caractère inintelligible et par conséquent son « acceptabilité globale » ne sauraient être contestés par des enseignants de langue.

7.2 Les liens de causalité avec la langue maternelle

Comme nous l'avons soutenu dans la section 1.3.2 certaines approches linguistiques avancent l'argument selon lequel l'utilisation d'une langue étrangère est caractérisée ou conditionnée dans une large mesure par une influence à la fois consciente ou inconsciente de la langue maternelle. De plus, selon un des tenants majeurs de cette approche dite de l'analyse contrastive, ce « conditionnement » s'opère indifféremment du niveau de maîtrise de l'apprenant. A titre d'illustration, dans les stades précoces ou intermédiaires de l'acquisition, l'on s'attend à ce que l'influence de la langue maternelle se manifeste bien distinctement dans les productions orales et écrites : à l'oral, il serait question d'un transfert du système phonologique intériorisé de la langue source qui provoquera, entre autres, un effet d'« accent étranger », tandis qu'à l'écrit les particularités lexico-grammaticales pourraient conduire à des phrases incompréhensibles ou agrammaticales.

Mais comme nous le verrons dans les sections qui suivent, nos résultats ne permettent en l'état ni de corroborer ni de réfuter complètement les hypothèses de transfert ou d'interférence. Cela est principalement dû au fait qu'aucune différence majeure n'est apparue entre ce qui est considéré comme de possibles erreurs de transfert entre les trois profils linguistiques de nos participants. Et compte tenu des différents contacts linguistiques existant entre ces derniers, un écart statistiquement significatif aurait permis de mieux délimiter les notions d'interférence, ce qui ne s'est pas produit. De surcroit, sans nier les transferts effectués par nos apprenants (dans les items étiquetés dans le premier volet), nous soulignons le fait qu'identifier une erreur comme étant un phénomène d'interférence revêt un caractère hautement subjectif de la part de l'analyste, dans la mesure où ce processus requiert que ce dernier ait connaissance de l'ensemble des répertoires linguistiques auxquels tous les apprenants auraient accès – ce qui est difficilement possible à en croire les nombreuses disparités linguistiques identifiées chez les sujets-participants et mises en avant dans l'annexe A2. L'analyste doit donc être conscient de cette limite, notamment s'il souhaite être impartial et appliquer la même systématicité partout dans son analyse.

7.2.1 L'influence de la langue française sur l'anglais

Certains étiquetages issus du schéma d'annotation d'UAM (volet 1) portent sur l'influence de la langue maternelle des apprenants et tout particulièrement sur les différents éléments qui ont été transposés en anglais langue étrangère, en termes d'interférence ou de transfert. Ces étiquetages incluent, entre autres, les erreurs d'emprunt, de calque, ou voire d'éléments phraséologiques (cf.

notamment les sections 5.1.1 et 6.1.2). L'ensemble de ces « transferts » fait l'objet d'une comparaison dans cette sous-section afin (i) de juger du niveau d'interférence de la langue française dans la rédaction en anglais (L2) parmi nos trois groupes tests et (ii) d'identifier les éléments linguistiques qui « subissent » le plus de transfert dans ces trois groupes. Les tableaux 40 et 41 présentent le résultat de ces croisements synthétiques. Les chiffres renvoient à la fréquence individuelle observée par item et par apprenant.

LEXICAL-ERROR-TYPE	E+CLR	F-CLR	F+CLR
spelling-error	14,6	9,4	10
vocab-choice-error	2,2	2	1,2
false-friend	0,6	0,6	0,4
other-wordchoice-error	1,8	2,2	1,2
coinage	0,4	2	1
borrowing	1,2	0,4	0,2

Tableau 40 : Regard croisé sur les transferts lexicaux

Dans le tableau 40, l'orthographe a été retenue comme « erreur de transfert ». Une remarque toutefois s'impose : notons la différence d'écart d'une part entre les participants étrangers (E+CLR) et les participants francophones (F-CLR et F+CLR) d'autre part. Cet écart signifie que l'orthographe est un problème particulièrement épineux pour l'ensemble des trois groupes – mais qu'elle se manifeste de façon très marquée chez le premier groupe. Cela étant dit, malgré le faible effectif illustré dans le tableau, rappelons que les erreurs lexicales regroupent un total de pas moins de 1007 occurrences annotées dont 654 appartenant à la catégorie des erreurs d'orthographe.

Pour ce qui est donc des autres écarts observés, signalons que le groupe F+CLR qui a bénéficié d'un contact linguistique prolongé en dehors du cadre institutionnel recense le moins d'items annotés dans chacune des six catégories citées. Ce constat est donc sans appel. Mais sans passer en revue tous les autres éléments du tableau, ce qui retient notre attention tout particulièrement ici est le niveau d'écart entre l'emploi des « false-friends », « coinage » et « borrowing[114] ». Pour le premier cas, l'écart observé est assez minime, ce qui laisse penser que l'emploi des termes appartenant à nos deux langues d'études mais dont le sens n'est pas identique pose le même niveau de difficulté aux trois sous-groupes. Pour ce qui est du deuxième cas (*coinage*, c'est-à-dire des emprunts de la langue française adaptés à la morphosyntaxe de la langue anglaise), on remarque que les participants F-CLR et F+CLR en font plus que les participants E+CLR tandis que l'inverse se réalise avec les erreurs de *borrowing* (à savoir des calques ou emprunts de la langue française non-adaptés à la morphosyntaxe de la langue anglaise). Les participants « francophones » auraient

[114] Cf. section 6.1.1 pour un rappel succinct des différences entre ces termes et des exemples correspondants.

donc deux à quatre fois plus tendance à incorporer un élément de leur langue maternelle en essayant de l'adapter linguistiquement à l'anglais, tandis qu'à l'inverse les participants E+CLR auraient trois à six fois plus tendance à emprunter des termes du français sans les adapter en anglais.

Ainsi, les participants francophones semblent reconnaitre à la fois une certaine similarité de certains termes français et anglais et aussi une différence dans ces deux langues dans la mesure où ils cherchent également à adapter ces « emprunts » de la première langue vers la deuxième. Les E+CLR, en ne modifiant pas les « emprunts » semblent dans une certaine mesure ignorer ces différences. Et ce refus de reconnaitre les différences mène au constat dans lequel on note une transposition très appuyée du lexique du français langue étrangère vers l'anglais qui est la troisième (ou quatrième) langue étrangère chez certains participants E+CLR. Ce constat inattendu rejoint une tendance déjà mise en avant, entre autres, par Tremblay (2006) et Forsyth (2014). En effet, selon l'étude de Tremblay (2006) il semblerait que la première langue étrangère (L2) ait une plus grande influence sur la L3 dès lors que les apprenants ont un contact très renforcé avec la L2. Ce cas de figure est bien identique à nos participants E+CLR qui ont des langues maternelles différentes, mais ont tous le français comme première langue étrangère (L2) et l'anglais comme deuxième ou troisième langue étrangère (L3). Cela étant dit, les taux de transferts observés dans les erreurs de « borrowing » et également d'orthographe (cf. la section suivante) suggèrent que ces participants considèrent le français plus proche de l'anglais que de leur langue maternelle. Passons maintenant aux éléments de transfert repérés au niveau proprement phraséologique.

PHRASING-ERROR-TYPE	E+CLR	F-CLR	F+CLR
transferred-phrasing	0,8	2,8	1,2
other-phrasing-error	3,4	1,8	1,4
phraseology-error	2,2	1,8	0,8

Tableau 41 : Regard croisé sur les transferts phraséologiques

Pour ce qui a été étiqueté dans le schéma d'UAM comme étant des erreurs de transfert, le groupe F-CLR semble le plus enclin à procéder aux transferts des unités multi-mots du français vers l'anglais. Ces unités signalées en section 6.1.4 renvoient aux expressions, souvent idiomatiques, transposées et traduites telles quelles en anglais : par exemple « dans un premier temps » *'*in a first time'* ou « avoir pour but » *'*to have for aim'*. Ce procédé constitue alors un problème considérable pour le groupe F-CLR, le groupe F+CLR et dans une moindre mesure pour les E+CLR. Pour ces derniers, ce manque de transfert à ce niveau peut s'expliquer par le fait que ces participants n'ont pas encore intériorisé ces unités multi-mots en français et ne peuvent donc pas encore les transposer

en anglais. De même, cependant, les erreurs diverses de mise en texte ou de phraséologie sont plus nombreuses chez les E+CLR. Ces erreurs renvoient à des unités vraisemblablement fonctionnant comme des ensembles mais que l'analyste n'a pas pu ranger dans les catégories d'erreurs phraséologiques existantes. Au vu de ce tableau, il semble en effet qu'il s'agira d'éléments transférés de la L1 de ces participants, ce qui expliquerait l'étiquetage attribué et la fréquence observée dans ce groupe par rapport aux deux autres. Toutefois ces propos sont à nuancer étant donné l'effectif observé chez ces derniers (F-CLR et F+CLR) et qui ne peuvent pas non plus renvoyer à des possibles transferts de la langue française.

Au vu donc de ces constats mitigés deux possibilités d'explication s'offrent à nous : il est possible soit de considérer ces erreurs comme relevant des aléas de toute rédaction soit de les voir comme une tentative de « créativité » de la part des apprenants qui chercheraient à mettre en place ou utiliser un effet de style qui n'a pas donné le résultat escompté en langue étrangère. Quant aux items phraséologiques restants, force est de constater qu'une fois de plus l'écart est très minime entre E+CLR et F-CLR, mais beaucoup plus entre ces deux derniers et F+CLR. Ces erreurs majoritairement de défigement phraséologique – qui rappelons-le, relèvent d'un procédé similaire aux simples erreurs de transferts – se différencient ici par le fait que les apprenants n'ont pas seulement traduit des items phraséologiques, mais les ont « défigé » en les « adaptant » à la langue anglaise. Par exemple « au contraire » *'*in contrary'* ou « être en vue » '**to be in the sight'*.

En définitive, les erreurs de transferts n'ont pas à première vue paru très statistiquement significatives, étant donné les faibles effectifs et les écarts qui varient d'un point ou deux. Elles ont tout de même réussi à indiquer des tendances qui semblent corroborer les études antérieures, notamment pour ce qui est des transferts d'erreurs lexicales d'une langue étrangère à une autre. De plus, si l'on considère que les F+CLR ont enregistré les plus faibles effectifs observés dans l'ensemble des catégories comparées dans cette section, ce constat constitue un point qui confirme une hypothèse de Van de Craats (2002). En effet selon cette dernière, « [t]ransfer seems also to be dependent on the nature of exposure to L2. Hence, it occurs more frequently in the beginning of the acquisition process, and when L2 is learned in an L1 environment (e.g., in schools and in situation where learners have little contact with L2 speakers... » Cela explique en partie pourquoi les F+CLR qui ont eu un contact prolongé avec l'anglais en dehors du contact institutionnalisé de leur L1 semblent avoir moins d'éléments transférés dans leur rédactions en L2.

7.2.2 Les limites de la notion de transfert et d'interférence

En considérant les différentes catégories d'erreurs telles qu'établies par le schéma d'annotation d'UAM, il est possible d'identifier des items « un peu partout » qui s'apparenteraient à des erreurs de transfert. C'est-à-dire qu'il est possible de repérer et de déduire une sorte de pseudo-lien ou « résidu » d'une langue source dans l'ensemble des occurrences étiquetées. Ces rapprochements semblent, à notre sens, cacher une analyse fortuite, mais recouvrent en réalité un phénomène bien plus enraciné dans le processus acquisitionnel d'une langue étrangère qu'on ne le pense. En effet, en section 5.1.1, nous avons mis en avant le fait que la quasi-totalité des 1007 erreurs lexicales s'était produite de manière non-systématique – et plus précisément avec un niveau de récurrence des items individuels très faible vis-à-vis des 122 sujets-participants de l'étude. Mais si l'on souhaite procéder à une sorte de rapprochement forcé, chacune des différentes erreurs lexicales peut être expliquée à la lumière d'un possible transfert entre langue source et langue cible. Cette possibilité est illustrée dans le tableau ci-après.

source (schéma UAM)		**exemple**
lexical	spelling (a)	i. worste ii. methode iii. controle
	spelling (b)	iv. developpement v. emmissions vi. investissement
	spelling (c)	vii. theorically viii. theoric
	spelling (d)	ix. compagny x. denonce xi. exemple xii. objectif
	spelling (e)	xiii. ingeniering
	false-friend (f)	xiv. education
	transfer (g)	xv. changement
	transfer (h)	xvi. to acceed xvii. discuted xviii. applicate xix. explication

Tableau 42 : Quelques exemples de transfert dans les erreurs lexicales

Dans le tableau 42, il devient alors possible de considérer tous les items comme un problème de transfert, comme exemplifiés dans les sept cas de figure suivants.

(v) spelling (a) : On pourrait mettre en avant le fait que les consonnes finales sont en principe muettes en français, sauf quand elles sont suivies par une voyelle. Si l'on suppose donc que les valeurs phonétiques des différents items lexicaux sont connus par les apprenants (peut-être les ont-il déjà entendus en classe ou ailleurs), ces derniers vont, de ce fait, les écrire avec un <e> final selon la convention ou l'orthographe française – permettant ainsi de bien indiquer la présence des sons /t/, /d/ et /l/ et d'éviter par la même occasion des consonnes muettes si l'on doit aller jusqu'au bout du raisonnement.

(vi) spelling (b) : Le dédoublement des consonnes étant a priori bien plus fréquent en français qu'en anglais, on pourrait donc conclure ici à un simple transfert phonie-graphique de la première langue vers la seconde. Il y aurait donc ici un transfert à la fois à l'écrit et à l'oral.

(vii) spelling (c) : Les deux orthographes pourraient s'expliquer de deux manières : (i) soit l'apprenant écrit ce qu'il pense avoir entendu (dans quel cas ce n'est plus un transfert) ; (ii) soit ce dernier procède au transfert de l'adverbe « théoriquement » en français en remplaçant le suffixe correspondant à la « fonction adverbiale » par son équivalent en anglais. Dans ce deuxième cas, il s'agirait effectivement de transfert. Hélas, nous n'avons pas les moyens de vérifier le processus cognitif qui a conduit au choix final et ne pouvons donc privilégier aucune des deux hypothèses.

(viii) spelling (d) : Il s'agirait ici en principe d'un phénomène de transfert par excellence dans la mesure où les deux formes existent de manière quasi-identique en L1 et en L2 avec la même signification. Le transfert se manifestera donc dans le fait que l'apprenant opte de façon inconsciente pour la forme et l'orthographe qu'il a l'habitude d'employer.

(ix) spelling (e) : Ce cas sera également un exemple de transfert par excellence dans la mesure où il est question d'un effet d'adaptation de l'orthographe proprement française à la prononciation du mot en anglais.

(x) False-friend (f) : Si l'on se réfère à l'exemple donné en section 5.1.1[115], ce terme a été employé un anglais avec transfert sémantique de la langue française.

(xi) transfer (g) et (h) : Ensemble, ces deux exemples renvoient de la même manière que « spelling (e) » aux tentatives d'adaptation des items lexicaux issus de la langue française que les apprenants cherchent à utiliser en les adoptant à la morphologie de

[115] Cf. 5.1.1 et plus précisément la section « False-friend errors » pour une discussion plus approfondie.

la langue anglaise[116]. Pour « *transfer (g)* », il n'y a pas de modification et pour « *transfer (h)* » il y a effectivement une tentative d'adaptation. Dans ces deux cas, on est véritablement ici dans le sens traditionnel du processus de transfert.

À la lumière de ces exemples, il en ressort qu'il est possible de procéder à un rapprochement entre la notion de transfert et tous les items annotés en tant qu'erreurs lexicales. Mais, il est également possible de faire de même avec les erreurs dites de grammaire, de pragmatique voire l'effet de la ponctuation sur la phraséologie (cf. section 8.2.2.2) Toutefois, bien que l'analyse semble aller dans le sens d'un transfert avéré, notre expérience en tant qu'enseignant de langue étrangère et en tant qu'apprenant de plusieurs langues étrangères nous oblige à tempérer cette analyse qui nous parait surprenante. Il nous semble en effet que l'ensemble des erreurs observées ne peuvent pas relever d'un même phénomène de transfert. De plus, s'il s'agit effectivement de transfert, cela ne se présente pas de la même manière à tous les niveaux d'analyse que nous avons passés en revue dans la présente étude.

7.2.3 Vers la réhabilitation de l'interlangue

Au vu donc des écarts qui n'ont été ni aussi signifiants qu'escomptés, ni aussi distincts les uns des autres en ce qui concerne nos trois groupes tests, nous émettons l'hypothèse que certaines erreurs témoignent plutôt d'une sorte de reconfiguration procédurale que nous ne pouvons ni expliquer intégralement en termes d'écart par rapport à la langue source ni expliquer totalement parce que les occurrences ne sont pas « en parfaite adéquation » avec la langue cible. Cette reconfiguration reflète, à notre sens, un système linguistique à la fois embryonnaire et intermédiaire composé de différentes représentations et connaissances acquises et apprises qui auraient été intériorisées par l'apprenant. Nous rejoignons ici d'une certaine manière la notion d'interlangue de Selinker (1972). Cependant, nous pensons que l'on gagnerait à approfondir et délimiter davantage les notions de transfert en identifiant d'une part, ce qui relèverait proprement d'un transfert du système linguistique de la langue maternelle que l'apprenant aurait acquise et d'autre part, ce qui relèverait d'un système de connaissances institutionnalisées que l'apprenant aurait apprises. Il pourrait également être question de faire une distinction entre les transferts réellement identifiables au niveau lexico-grammatical, discursif, sémantique dans le but d'étudier les différentes facettes de la notion complexe, mais chacune selon un angle différent. Toutefois nous sommes conscient des limites de ce que nous proposons : (i) notamment en raison du fait que l'hypothèse de

[116] Nous renvoyons à nouveau à la section 5.1.1 aux sections « Coinage » et « Borrowing ».

reconfiguration nécessiterait une analyse psycholinguistique approfondie si l'on souhaite obtenir des précisions sur ce phénomène de reconfiguration ou de connaissances amalgamées et (ii) délimiter les pratiques linguistiques de l'apprenant entre ce qui relève du système lui-même et ce qui relève de l'acculturation ne sera pas chose aisée.

7.3 Bilan de l'influence du temps et des différentes rencontres linguistiques

En réexaminant les erreurs relevées dans les chapitres V (volet 1) et VI (volet 2), nous avons cherché à étudier les différents « comportements » d'une des variables majeures de la présente étude – dans les trois sous-groupes linguistiques identifiés parmi nos 122 sujets-participants. Cette variable temporelle avait pour but de mettre en avant les écarts de performance relevés dans nos deux semestres d'observation chez les trois groupes tests aussi bien dans les erreurs du système linguistique (volet 1) que les erreurs d'acceptabilité textuelle (volet 2). Plusieurs analyses ont été faites par rapport aux débuts précoces de certains apprenants ou une initiation tardive dans l'apprentissage de l'anglais chez les autres. Les différentes rencontres linguistiques chez les participants ont également été prises en compte.

En définitive il est en ressorti que (i) peu de différences ont été soulignées entre les participants ayant commencé l'apprentissage de l'anglais à la maternelle et ceux qui l'ont commencé au collège ; (ii) les apprenants qui avaient effectué des séjours prolongés dans les pays anglophones (F+CLR) continuent à faire autant d'erreurs d'un semestre à l'autre voire davantage dans certains cas, et ce, contrairement aux groupes de participants francophones sans séjours en pays anglophones qui ont tous enregistré des baisses dans le deuxième semestre par rapport au premier. Mais ce surcroit d'erreur dans le groupe F+CLR semble traduire, à notre sens, de multiples prises de risque et davantage d'implication personnelle lors des productions écrites tandis que les autres candidats se contentent plutôt d'utiliser leurs acquis jugés stables, sans trop vouloir sortir de leur « zone de confort ».

Il est également à rappeler que les erreurs relevant du système linguistique (volet 1) ont diminué de plus de 80% chez l'ensemble des participants-tests, tandis que les erreurs textuelles n'ont baissé que de 47%. Cet écart semble confirmer le postulat que les deux types d'erreurs ne sont pas régis de la même manière. Et tandis qu'un cours d'anglais de spécialité semble avoir un impact considérable sur le premier type d'erreur, l'impact n'est pas encore « au point fixe » pour ce qui est des erreurs textuelles. Cela étant, l'augmentation de certaines erreurs au deuxième semestre qui affectent tout particulièrement la structuration informationnelle indique que ces éléments méritent qu'on leur

accorde davantage d'attention en cours de langue : ne serait-ce que dans le but de freiner la progression ou d'anticiper davantage ces types d'erreurs et ainsi mieux les corriger.

Enfin, nous avons relevé des risques ou des limites liés à la classification d'éléments comme étant des transferts d'une langue à une autre, dans la mesure où le terme « transfert » semble pouvoir à la fois tout englober et ne rien préciser. Ce terme gagnerait alors à être mieux défini par rapport, bien entendu, à la langue source et la langue cible mais également la langue en construction chez l'apprenant, à savoir l'interlangue.

(Chapitre VIII) Conclusion : typologie des erreurs revisitée et perspectives

Ce dernier chapitre est divisé en quatre sous-sections distinctes. La section 8.1 rappelle tout d'abord le degré de validité que l'on peut accorder à l'ensemble de nos annotations, avant de résumer les principaux résultats obtenus dans la présente étude. Elle retrace ensuite les principales fréquences d'erreurs observées dans nos deux volets d'annotations, présentés respectivement dans les chapitres V et VI. La section 8.2 met en avant les observations les plus notables : dans la section 8.2.1 quelques aspects cognitifs sous-jacents à certaines erreurs seront expliqués ; et dans la section 8.2.2 nous nous pencherons sur les catégories d'erreur relevées au niveau de la mise en phrase, qui – selon nous – sont souvent insuffisamment prises en compte dans la littérature d'analyse d'erreurs traditionnelle. Dans la section 8.3, les principales observations de la présente étude seront comparées avec des travaux qui ont été menés dans un cadre similaire : (i) l'interface entre ce que l'on considère comme étant des erreurs lexicales, syntaxiques et sémantiques sera explorée ; (ii) le flou conceptuel et terminologique de certains étiquetages d'erreur, y compris ceux utilisés notamment dans le chapitre V, sera également discuté et (iii) un nouveau cadre d'analyse sera proposé. Enfin, la section 8.4 met en avant des questions périphériques qui ont accompagné notre réflexion tout au long de ce travail (cf. 8.4.1) et des limites que nous avons pu y identifier (cf. 8.4.2).

8.1 Synthèse des principaux résultats et leurs implications didactiques

Il convient de rappeler tout d'abord que l'ensemble de nos annotations sont généralisables à toute population d'étudiants ayant des traits similaires aux sujets-participants de notre étude. En effet, les différents scores de Kappa (cf. section 4.3) – qui oscillent entre 0.82 et 0.89, et qui ont été obtenus grâce à nos quatre annotateurs indépendants – ont démontré que malgré les différences entre les profils professionnels des annotateurs, les annotations obtenues sont sensiblement comparables à celles que nous avons obtenues. Le bien-fondé et la solidité des annotations et par voie de conséquence des résultats, présentés tout au long de ce travail, traduisent, à notre sens, un certain degré de fiabilité générale. A ce titre, nous pensons non seulement que nos résultats sont « généralisables » à d'autres étudiants francophones apprenant l'anglais, dans un cadre identique au nôtre, mais également que l'on risque d'arriver à des données statistiques similaires, si l'on reproduit la présente étude avec les mêmes paramètres, dans un autre établissement supérieur français. Passons maintenant à la synthèse des principaux résultats.

8.1.1 Volet 1 (Les erreurs du système linguistique)

Rappelons ici que l'étude a été divisée en deux volets d'annotations distincts. Le premier volet porte sur les erreurs dites du système linguistique et celles-ci conduisent inévitablement à des agrammaticalités ; le deuxième volet porte sur des erreurs d'acceptabilité textuelle et ces dernières nécessitent de manière générale un contexte plus large (au-delà de la phrase isolée) pour les apprécier à leur juste valeur. Le tableau ci-dessous fait état des quatorze points qui posent le plus de difficultés à l'ensemble des sujets-participants de notre étude, dans le premier volet d'annotation.

	Types d'erreurs	n	(%)
1	orthographe	653	12
2	choix-du-verbe-lexical	364	6,7
3	segment (phrastique) incomplet	329	6,1
4	déterminant présent non-requis	278	5,1
5	accord en nombre	250	4,6
6	déterminant absent-mais-requis	222	4,1
7	choix-de-préposition	207	3,8
8	accord-sujet-verbe	194	3,6
9	incohérence temporelle (clause complex)	185	3,4
10	saillance et incompatibilité des référents	181	3,3
11	divers-erreurs-de-mise-en-phrase	139	2,6
12	position adverbiale (ou d'ajout)	120	2,2
13	phraséologie lexicale	111	2,1

14	transfert phraséologique	108	2
	total[117]	*3341*	*62*

Tableau 43 : Les erreurs les plus fréquentes dans notre corpus (volet 1)

Deux remarques s'imposent ici. D'abord, les faibles pourcentages sont à mettre en rapport avec le fait que le schéma d'annotation – que nous avons utilisé pour étiqueter nos erreurs – comporte plus de 160 étiquetages différents. Il convient donc de comprendre qu'une telle granularité fait que les catégories sont assez précises, ce qui dans notre contexte permet d'identifier les points qui constituent de véritables écueils. La deuxième remarque est plus générale. Comme nous verrons dans la section 8.3, si l'on pense aux différentes études que nous avons comparées à la nôtre, il devient alors clair que les trois premières catégories du tableau 43 (c'est-à-dire, orthographe, choix-du-verbe et segment (phrastique) incomplet) ne figurent pas de manière générale dans les classements d'erreurs les plus fréquentes : soit parce qu'elles ne sont pas jugées importantes soit parce qu'elles n'ont vraisemblablement pas constitué des « étiquetages » à part entière dans ces différentes études. A cette remarque s'ajoute le fait que si l'on « enlève » ces quatre premières catégories d'erreurs, on retrouve un classement avec des éléments quasi-identiques tels qu'ils sont mis en avant dans de nombreux projets d'analyses d'erreurs que nous avons pu consulter (cf. par exemple Chuang & Nesi, 2006).

Au vu donc du tableau 43, une des conclusions que l'on peut en déduire est que la majorité des sous-catégories (numérotées de 4 à 14, et tout particulièrement 4 à 8) ne sont pas propres aux étudiants français apprenant l'anglais. Et contrairement à ce que Chuang & Nesi ont affirmé, ces erreurs ne sont pas non plus imputables à une influence de la langue chinoise[118]. Nous rappelons, de ce fait, que ces erreurs revêtent, à notre sens, un caractère inhérent à l'acquisition de l'anglais langue étrangère. En effet, les problèmes, par exemple, de déterminant, d'accord en nombre et de préposition sont parmi les plus cités dans la littérature[119]. On a donc affaire à des obstacles pour lesquels les enseignants doivent par conséquent « s'armer », puisque ces écueils « ont vraisemblablement la vie dure ». Ainsi, il nous parait primordial que les enseignants ne sous-estiment pas ces problèmes et qu'ils leur accordent une « attention corrective considérable » en cours de langue. Si cela n'est pas fait, les apprenants risquent de poursuivre leur apprentissage de l'anglais tout en commettant un certain nombre d'erreurs « très basiques » (cf. le tableau 43) *ad infinitum,* au point que certaines erreurs risquent de se fossiliser. Ce qui semble, à titre d'exemple, être le cas pour les différentes erreurs de déterminant chez un grand nombre de nos apprenants.

[117] Rappelons qu'uniquement quelques exemples ont été choisis ici, ce qui fait que le total ne fait pas 100%.
[118] Les apprenants de l'étude de Chuang & Nesi (2006) étaient de langue maternelle chinoise.
[119] Cf. par exemple Bitchener et al. (2005) à propos des erreurs de préposition.

8.1.2 Volet 2 (Les erreurs d'acceptabilité textuelle)

Dans le deuxième volet d'annotation, les erreurs de mise en phrase – qui renvoie dans une certaine mesure à des erreurs de phraséologie lexicale (cf. les sections 6.1.2, 6.2.1 et 8.2.2.1) – sont les plus fréquentes. C'est-à-dire lorsque la grammaticalité de l'énoncé n'est plus la source de l'erreur, il y a une probabilité de 25% que l'erreur vienne d'une sorte d'« incompatibilité phraséologique ». Il n'est donc pas un hasard que la deuxième erreur sur le classement dans ce volet relève du choix sémantique. Rappelons à titre d'information que les erreurs sémantiques sont souvent liées à des erreurs de choix collocationnels (cf. 6.2.1 et 8.2.2.1). Ce qui signifie que 48% des erreurs non-grammaticales sont imputables aux choix phraséologiques.

	n	(%)
E. de mise en phrase	175	25,58
E. sémantique	160	23,39
E. de progression temporelle *(type de focus 1)*[120]	80	11,69
E. référentielle	71	10,38
E. de coordination	55	8,04
E. d'agencement	50	7,30
E. de cadrage	36	5,26
E. d'ostension *(type de focus 2)*	35	5,11
E. de progression thématique *(type de focus 3)*	22	3,21
total	*684*	*100*

Tableau 44 : La fréquence des erreurs d'acceptabilité textuelle (volet 2)

Les erreurs de progression temporelle et les erreurs référentielles sont à mettre en rapport avec ce que nous avons appelé les erreurs de calcul sémantique (cf. section 8.2.1). Ce problème renvoie au fait que l'apprenant n'a pas correctement transféré l'ensemble des valeurs sémantiques et formelles accumulées sur les unités précédentes sur celles qui doivent, en principe, recevoir la marque de l'accord des opérations linguistiques qui ont précédé. L'erreur est, de ce fait, une erreur de forme, puisque l'apprenant mise de manière générale sur la continuité ou la construction sémantique au détriment de la correction formelle[121].

Les erreurs de coordination et d'ostension sont à mettre respectivement en rapport avec des constructions paratactiques et hypotactiques erronées[122]. Les erreurs de coordination sont principalement de types asyndétiques, c'est-à-dire qu'il manque l'élément clé permettant de coordonner les items énumérés ou juxtaposés par l'apprenant. Dans le cas des erreurs d'ostension,

[120] Rappelons que ces exemples renvoient aux trois sous-types d'erreurs que nous appelons des erreurs de focus.
[121] Notons que nous avons également observé le cas contraire, mais avec une fréquence moins élevée.
[122] Ces constructions sont réexpliquées en détail dans la section 8.2.2.2.

le lien entre l'exemple et la proposition à laquelle ce dernier est relié n'est pas suffisamment clair ou saillant – ce qui fait que la hiérarchisation ou la subordination attendue entre énoncés ne peut pas avoir lieu.

8.1.3 Synthèse

Par ailleurs, hormis la distinction entre les volets d'annotation et le cadre LSF, les résultats obtenus nous ont permis d'avoir une vue globale sur les manquements de nos apprenants – à plusieurs niveaux de leur production écrite. Les différentes catégories initiales du schéma d'UAM, y compris celles que nous avons ajoutées au fil de notre annotation, se sont toutes révélées probantes. Toutefois, notons que l'analyse faite au début de l'exercice d'annotation s'est avérée tout à fait différente de ce à quoi nous nous attendions, notamment pour la catégorie lexicale.

En effet, au vu des étiquetages initiaux, on pouvait penser dans un premier temps que les erreurs étiquetées « vocabulaire » étaient non-systématiques et donc aléatoires. Mais nous nous sommes vite aperçu que ça n'était pas le cas. Les erreurs lexicales étaient dues, en grande partie, à une non-conformité au niveau local de la phrase : c'est-à-dire que le choix lexical lui-même posait un problème d'acceptabilité par rapport aux autres items lexicaux disponibles dans la phrase. Cette erreur s'est révélée plus complexe que nous ne le pensions et nous soutenons, par conséquent, qu'elle ne sera pas simple à corriger en classe de langue. En effet, l'apprentissage par liste de mot ne suffira pas dans ce cas précis. Mais une meilleure prise en compte du fait que certains items entretiennent des rapports phraséologiques avec d'autres pourrait permettre de corriger le problème à sa racine, en aidant les apprenants à prendre connaissance du phénomène. Ces derniers pourraient par conséquent mieux anticiper et mieux identifier les associations qui existent, entre autres, par exemple, entre des items proprement phraséologiques.

S'ensuit alors la question du flou terminologique. En effet, étant donné que nous avons mis en avant le flou terminologique manifeste dans les différents schémas d'annotations que nous avons parcourus, il nous a semblé judicieux de ne pas en établir une nouvelle taxonomie d'erreurs ou un autre schéma d'annotation simplifié ou plus exhaustif : et ce, de façon à ne pas ajouter à la cacophonie ambiante. Cela dit, comme nous l'avons également démontré, bien qu'il y ait des différences au niveau de l'appellation présente dans les différents schémas d'annotations, les erreurs du même type (c'est-à-dire, qui sont directement transposables d'une étude à une autre) demeurent comparables. De plus, notre comparaison (cf. section 8.3.1.2, 8.3.2) indique que certains étiquetages permettent d'arriver à des conclusions similaires.

Nous militons pour l'adaptation des schémas existants – en non pour leur création *ad infinitum* – afin de faciliter des comparaisons plus exhaustives et d'éviter que l'analyste qui utilise un autre schéma d'annotation soit obligé d'isoler un faible nombre d'étiquetages parmi un schéma très exhaustif. Nous reconnaissons tout de même que c'est un peu idéaliste, mais nous pensons que l'emploi d'un schéma similaire entre plusieurs projets d'analyse d'erreurs ne pourrait qu'augmenter la validité d'un schéma donné, tout en permettant de recueillir davantage de données à mettre au profit des études en acquisition. De cette manière, l'analyste est sûr de pouvoir procéder objectivement à de véritables comparaisons.

Enfin, nous nous permettons une remarque d'ordre général. Il semblerait que les cours de spécialité (c'est-à-dire, non axés sur la grammaire, à proprement parler) ont permis une réduction considérable des erreurs grammaticales chez nos sujets-participants. Toutefois, ces cours n'ont pas eu le même effet sur les erreurs textuelles qui, rappelons-le, ne renvoient aucunement à une non-maîtrise linguistique. Il faudra donc accorder une attention particulière à ce problème, si l'on souhaite le corriger de manière objective en cours de langue. Reconnaissons tout de même qu'il faudra confirmer cette tendance, dans un futur projet, en comparant la progression de nos deux types d'erreurs, dans des cours d'anglais dits par exemple « de spécialité » (anglais des affaires, …) et un autre axé principalement sur la grammaire. Et (pourquoi pas ?) avec le même schéma d'annotation pour assurer « un bon degré de comparabilité » entre les différents étiquetages.

8.2 Quelques observations notables

Nous retraçons dans cette section quelques unes des observations qui nous ont le plus surpris de par : leur fréquence, leur caractère inattendu et le fait d'avoir des traits partagés entre erreurs de types différents.

8.2.1 Les problèmes de calculs sémantiques

Après avoir annoté et examiné plusieurs centaines d'erreurs de types différents, on commence à discerner non seulement les différences subtiles des balises d'erreur ou leur niveau de solidité mais, de façon intuitive, l'explication possible des erreurs observées. A ce titre, la plupart des erreurs s'avèrent promptement classifiables comme étant le résultat soit d'un manque de maîtrise linguistique – c'est-à-dire que l'apprenant n'a pas connaissance de la règle spécifique qui devrait être appliquée dans une situation donnée, soit d'une interférence où l'on peut observer un certain transfert de moyens linguistiques à la disposition de l'apprenant en L1 vers la L2. Mais, nous avons

également observé des erreurs qui ne peuvent être placées dans aucune de ces deux catégories, en raison du fait, d'une part, que l'environnement textuel immédiat apporte la preuve que les connaissances linguistiques ne peuvent pas être mises en cause, et d'autre part qu'aucune interférence ne peut être identifiée au-delà de tout doute raisonnable. Dans ces cas, nous sommes confronté à plusieurs questions supplémentaires : par exemple, est-ce que ces observations relèvent d'erreurs humaines que nous pouvons appeler ponctuelles ou est-ce que nous avons affaire à des erreurs propres à l'individu qui les aurait commises ?

Suite aux nombreux exemples consultés, nous en sommes venu à la conclusion que ces erreurs reflètent une certaine inattention qui est présente dans le travail d'un grand nombre de participants. Cette inattention semble toutefois être un phénomène « provoqué » plutôt qu'un phénomène aléatoire. Dans cette optique, elle ne peut donc pas être qualifiée d'évènement ou aléa de parcours non-systématique simple. Ce constat général fait écho dans une certaine mesure à celui de Chuang & Nesi (2006) dans lequel ils affirment que « students were not always capable of avoiding mechanical errors when they had to deal with the organisation of ideas and linguistic features at the same time ». Par conséquent, nous soulignons que lorsqu'il y a des demandes concurrentes - de façon simultanée - sur la 'forme' et le 'sens' (ou des besoins informationnels par opposition aux besoins structurels), les apprenants ont tendance à prioriser le deuxième et le premier est donc sacrifié. Nous appelons ces types de phénomènes des erreurs de calcul sémantique, étant donné que les apprenants doivent garder à l'esprit l'ensemble des petites opérations effectuées à un niveau aussi bien inter que intra-phrastique pour ensuite remporter "la somme accumulée" sur l'élément ou la phrase suivant(e). Et c'est justement la somme totale qui fait défaut aux apprenants puisque, soit ils ne remportent qu'une partie de la somme accumulée soit qu'ils l'abandonnent complètement. Nous allons aborder ces phénomènes plus en détail dans les trois sous-sections ci-après.

8.2.1.1 La concordance des temps et des auxiliaires modaux

Les erreurs annotées en tant que « problème de temps grammatical » et « problème de modal » (cf. section 6.1.1) dans le premier volet d'annotation et en tant qu'erreur de concordance de temps dans le deuxième volet relèvent ostensiblement du même type d'occurrence. Il nous parait donc judicieux de les regrouper sous l'étiquette « erreurs de concordance ». Le résultat est illustré ci-après.

	UAM [1er][123]	UAM [2ème]
concordance des temps	*7,46%*	*10,41%*

Tableau 45: les erreurs de concordance (volet 1)

Le regroupement permet de mettre en avant ce problème qui jusque-là n'avait bénéficié que d'une attention singulière (cf. Kübler 1995). Il est donc important de souligner ici non seulement sa fréquence et sa distribution par rapport à l'ensemble des erreurs grammaticales dans le premier volet d'annotation (cf. chapitre V) mais également son classement dans le deuxième volet portant sur les erreurs d'acceptabilité textuelle (cf. chapitre VI).

	2ème volet
E. de mise en phrase	25%
E. sémantique	13,3%
E. de concordance des temps	11,69%
E. de référence	10,38%
total[124]	*60%*

Tableau 46: les erreurs les plus fréquentes dans le volet 2

Au vu du nombre total d'étiquetages proposé par le schéma d'annotation d'UAM, le taux de 10,41% (volet 1) et 11,69 % (volet 2) indique tout d'abord que les erreurs de concordance sont présentes aussi bien dans des phrases grammaticalement correctes que dans des phrases agrammaticales. Ensuite, elles représentent le troisième phénomène linguistique qui provoque le plus d'occurrences erronées (notamment dans le deuxième volet d'annotation, cf. chapitre VI). En effet, si aucune occurrence de ce type n'avait été relevée dans le deuxième volet d'annotation, on aurait pu déduire que la concordance de temps constituait un problème majeur uniquement pour ceux dont la maîtrise grammaticale ou la maturité syntaxique n'était pas encore consolidée. Toutefois, le même type d'erreurs se trouve également dans des phrases tout à fait acceptables, sur le plan grammatical – ce qui laisse entendre que le problème dépasse la notion de grammaticalité.

Nous pensons par conséquent que ce type d'erreurs reflète la relative complexité cognitive de certaines structures grammaticales : autrement dit, la concordance de temps demanderait plus d'attention « cognitive » que, par exemple, un accord entre un pré-modifieur et le nom qu'il modifie. Le raisonnement découlant de cette hypothèse vient des 185 occurrences sur un total de 232 qui ont été identifiées dans des phrases complexes, où le temps grammatical utilisé dans la proposition principale régit et limite les choix possibles dans la proposition dépendante. A cela

[123] UAM [1er] renvoie au pourcentage initial obtenu dans le volet 1. UAM [2ème] renvoie au pourcentage obtenu après une reclassification que nous avons été contraints à réaliser pour procéder à des comparaisons objectives avec d'autres études. (cf. notamment 8.2.2 pour plus de précisions)

[124] Rappelons qu'uniquement quelques exemples ont été choisis, à titre d'illustration, ce qui fait que le total ne fait pas 100%.

s'ajoute le fait que l'exigence qui vient du contexte textuel plus large voudrait que le temps choisi assure une fonction de liage et de cohérence entre l'ensemble des informations présentées. Cette hypothèse vaut tout aussi bien pour les auxiliaires modaux[125] qui ont été signalés comme étant textuellement inacceptables au vu du contexte à la fois textuel et temporel.

8.2.1.2 L'accord en nombre : SN tête (PR1 et PR2)

Un autre cas où il semble y avoir un certain « comptage » ou « calcul » sémantique incorrect concerne le syntagme nominal et les rapports privilégiés avec ses satellites à gauche et à droite. Ce problème se manifeste par excellence à travers la relation entretenue entre le déterminant et le n-tête. Dans ces cas précis, plusieurs occurrences ont été signalées dans lesquelles il n'y a pas d'accord syntaxique entre les deux éléments précités : c'est-à-dire, le déterminant est dans une forme soit plurielle soit singulière avec laquelle le n-tête est incompatible. Toutefois, contrairement à la section précédente où les erreurs de calcul sémantique ont été vues dans des phrases grammaticales et agrammaticales, le problème d'accord entre déterminant et son n-tête conduit systématiquement à une erreur syntaxique. Par ailleurs, un autre cas d'erreur de comptage sémantique apparaît dès lors que le contexte textuel immédiat (postérieur ou précédent) exige que le noyau nominal soit d'un certain nombre et que l'apprenant ne parvient pas à satisfaire convenablement cette exigence au niveau de la phrase locale. Deux exemples sont fournis ci-dessous.

99. Education is one of the **priority* $priorities$[…] (txt_027_sm2)
100. But would all the economy be ravaged by the death of advertising? Isn't there any market that doesn't require any kind of advertisement? There are millions of answers to **that *question* $these questions$. (txt_004_sm1)

Ce problème d'accord mérite lui aussi d'être intégré dans le classement. Nous choisissons donc de poursuivre la comparaison avec une étude dans laquelle ce problème a été classé deuxième sur la liste des points les plus épineux : à savoir celle de Chuang & Nesi (2006) que nous avons déjà évoquée dans la section précédente. Soulignons tout de même une réserve quant à la comparaison avec ces derniers. En effet, leur classement a été établi par rapport à l'ensemble des étiquetages disponibles dans leur schéma d'annotation et les erreurs annotées n'ont pas été regroupées et présentées par catégorie plus large – ce qui fait que nous ne pouvons pas objectivement comparer les pourcentages obtenus. Nous nous référons uniquement ici au rang attribué par ces derniers.

[125] Cf. section 6.2.3 pour une présentation graphique, à travers le schéma interpersonnel de la LSF, qui met en avant la répartition générale des erreurs de modal, d'auxiliaire et les autres erreurs portant sur le temps grammatical.

Par ailleurs, Chuang & Nesi ont choisi de présenter les items situés au dernier niveau de leur schéma d'annotation – c'est-à-dire, au lieu de signaler qu'on a tout d'abord affaire à une erreur dans le syntagme nominal, ensuite sur le déterminant et ainsi de suite, les erreurs sont uniquement présentées selon les dernières catégories établies sur les déterminants. Nous allons, à titre comparatif, faire de même sur les éléments présentés dans deux tableaux différents (cf. tableaux 45 et 54) de manière à mieux illustrer nos propos.

	UAM [2ème] %	Chuang & Nesi
déterminant-présent-mais-non-requis	12,48%	3ème (8,5%)
accord-en-nombre	11,22%	2ème (8,8%)
concordance des temps	10,42%	N/A
déterminant-absent-mais-requis	9,96%	1er (10,1%)

Tableau 47 : Classement comparatif de quatre éléments problématiques (volet 1)

Ce tableau permet de voir que les déterminants sont les plus problématiques dans les trois[126] études comparées, mais pas de la même manière selon qu'ils sont dans la nôtre ou celle de Chuang et Nesi. Ce qui nous intéresse tout particulièrement ici, cependant, ne sont pas les déterminants (cf. 5.1.2.1) à proprement parler mais les erreurs d'accord en nombre qui ont été classées deuxième en termes de fréquence dans les deux études.

	PR1	PR2	*total*
sm1	32	87	*119*
sm2	23	108	*131*
total	*55*	*195*	*250*

Tableau 48 : Répartition des problèmes d'accords en nombre

Et bien que nous n'ayons pas les moyens psycholinguistiques dans le présent travail de vérifier notre hypothèse qui voudrait que les accords en nombre traduisent une sorte d'inattention ou charge cognitive supplémentaire, il devient évident – à la vue du classement comparatif dans le tableau 5 et celui du tableau 6 que nous avons établi suivant le principe de PR1 et PR2 issu du schéma expérientiel (cf. chapitre V, section 5.2) – que le problème des accords en nombre survient vraisemblablement quand la « somme sémantique » du calcul mental effectué dans la première partie de la phrase n'a pas été transférée vers la deuxième partie. Le tableau 6 nous indique, en effet, sans surprise que 78% des erreurs d'accord en nombre se produisent dans le rôle de PR2, qui - pour rappel – renvoie aux syntagmes nominaux jouant le rôle de complément par opposition au

[126] A savoir celle de Chuang & Nesi (2006), celle de Dagneaux et al (1998) que nous abordons notamment dans la section 8.3.1.2 et la nôtre.

rôle de sujet grammatical dans une phrase donnée. Et c'est justement cette deuxième partie qui est supposée s'accommoder de manière générale de la somme des opérations syntaxiques qui l'a précédée.

8.2.1.3 Les problèmes des chaînes de référence

Le problème du calcul sémantique ne se limite ni aux erreurs de concordance des temps ni aux accords en nombre, il est tout particulièrement présent dans les erreurs de référence et/ou de coréférence. Cependant, peu d'études ayant des paramètres conceptuels et méthodologiques similaires aux nôtres s'y sont intéressées à ce point. En effet, la difficulté à laquelle on doit faire face ici se distingue de deux manières : soit il est question d'un problème de saillance entre référents ; soit il s'agit d'un problème d'incompatibilité totale entre référents (cf. chapitre VI, section 6.1.1 pour des exemples). Force est de constater par ailleurs que cet écueil peut conduire à la fois à des questions de grammaticalité (cf. section 6.1.1) et d'acceptabilité (cf. section 6.2.1), sans oublier bien entendu qu'il semble y avoir une prédilection pour une catégorie ou classe de mots en particulier : à savoir les pronoms, tous sous-types confondus.

Parmi les signalements d'erreurs de cohésion dans le chapitre VI, on pourrait préciser, entre autres, les problèmes d'incompatibilité des adjectifs possessifs (qui sont présents à hauteur de 19% dans le corpus) notamment dès lors que l'apprenant cherche à établir le lien de référence avec un possesseur donné ; à cela s'ajoute les inexactitudes dans les usages des pronoms démonstratifs (28%), les pronoms relatifs (2%), les pronoms réfléchis (1%) et les pronoms divers ayant un rôle de sujet (31%) et un rôle d'objet ou de complément (19%). Notons enfin que cette difficulté survient aussi bien entre des items dans une chaîne référentielle à courte, moyenne et longue distance. Autrement dit, à l'intérieur d'une même proposition, entre deux propositions d'une même phrase et entre deux phrases indépendantes. Trois exemples sont fournis pour illustrer ce propos.

101. Anyone in **its* his right mind would say the Union is doomed […]. (txt_004_sm2)
102. The poorest members are in a difficult position and the ones that are barely able to go through the crisis have to help them because **they* have consequences on **them* $##$. (txt_017_txt_sm2)
103. Moreover, social protection *prevent[sic]* them from hardships at work. They sometimes learn their rights and how legislation works at the university. **This one* $University$ also delivers new tools, useful on the labour market (for instance how to write a CV, a letter [...]. (txt_030_txt_sm2)

Ce qu'il faut retenir ici c'est que les moyens linguistiques qui ont pour fonction principale de se substituer à un antécédent peuvent faillir à leur rôle, dans la mesure où l'apprenant n'a pas gardé à l'esprit toutes les propriétés linguistiques des antécédents avec lesquels il cherche à établir un lien de référence. A ce titre, la somme de ces propriétés linguistiques n'est donc pas suffisamment prise en compte et ne peut, par conséquent, être correctement transférée. Cet oubli ou inattention conduit aux différents écueils soulignés ci-dessus : à savoir des liens insuffisamment saillants entre l'antécédent et le pronom qui le remplace ou une totale incompatibilité avec l'antécédent auquel un nouveau référent est supposément relié.

En définitive, la présence de l'ensemble de ces problèmes chez de nombreux participants de notre étude – que l'on pourrait par analogie comparer à des échecs dans les calculs mentaux ou dans des sortes de calculs sémantiques – témoigne de la nature non-aléatoire de ce phénomène. Nous réitérons donc notre postulat que cette erreur est une erreur "provoquée" plutôt qu'une erreur aléatoire qui survient lorsqu'il y a des exigences concurrentes - de façon simultanée - sur la 'forme' et le 'sens'. En effet, ce type d'erreur apparaît de manière générale quand il y a une demande intensive de précision imposée conjointement par le système linguistique lui-même et le texte dans son besoin de cohésion textuelle. Cette double exigence nécessite une certaine maturité syntaxique de la part de l'apprenant et les erreurs de calcul sémantique prennent forme lorsque l'apprenant est incapable de répondre de façon convenable à ces deux exigences. Ce dernier choisit donc ici de faire attention au « sens » au détriment de la régularité formelle.

8.2.2 Les erreurs de mise en phrase

Parmi les nombreuses erreurs que nous avons rencontrées dans le corpus, beaucoup dépassent les items lexicaux individuels simples et ont donc été annotées comme relevant de l'ordre de la phrase : c'est notamment le cas des erreurs signalées en section 6.1.2 et quelques unes par exemple, entre autres, dites de cadrage ou de mise en phrase expliquées dans la section 6.2.1. Mais après avoir regroupé des items annotés dans un premier temps comme étant principalement lexicaux, nous nous sommes aperçus que certains demeurent non seulement inacceptables au vu du contexte informationnel mais tout particulièrement au vu des rapports sémantiques et syntaxiques entretenus avec d'autres items lexicaux qui les entourent. Ce constat nous a conduit à effectuer des analyses supplémentaires.

Types d'erreurs	UAM [1er]		UAM [2ème]	
Er. Lexicale	1007	18,6	1887	34,9
Er. Grammaticale	3107	57,4	2227	41,2
Er. Pragmatique	385	7,1	385	7,1

Er. Phraséologique	687	12,7	687	12,7
Er. Incodable	22	0,4	22	0,4
Er. de ponctuation	202	3,7	202	3,7
total	*5410*	*100 (%)*	*5410*	*100 (%)*

Tableau 49: Un aperçu modifié de l'ensemble de nos résultats

En effet, en retenant la classification d'erreurs lexicales de Granger & Monfort (1994) et de Dagneaux et al. (1998), nous avons pu regrouper ensemble la totalité des choix lexicaux erronés préalablement identifiés dans le chapitre V. Rappelons, à titre d'information, qu'en suivant le schéma d'annotation d'UAM CorpusTool certaines 'erreurs lexicales' avaient été rangées dans diverses catégories grammaticales : par exemple, les erreurs lexicales sans incidence sur la grammaticalité de la phrase ont été classées selon la classe ou nature des mots mis en cause (c'est-à-dire, dans la section 5.1.2 nous avons « choix erroné du verbe lexical », « choix erroné de préposition », « choix erroné de nom », et ainsi de suite). Le résultat de ce regroupement est illustré dans le tableau ci-dessus, et l'on voit clairement que le rapport de force entre erreurs lexicales et erreurs grammaticales change de manière considérable entre UAM [1er] (les résultats avant modification) et UAM [2ème] (les résultats après modifications). En bref, ce que nous avons cherché par la suite dans ce regroupement c'est de savoir si les 880[127] nouvelles erreurs lexicales (ou les erreurs nouvellement catégorisées) sont réellement lexicales ou si elles relèvent d'un autre ordre. Ce que nous avons observé est détaillé dans la section suivante.

8.2.2.1 Les erreurs de phraséologie lexicale

Parmi les 880 erreurs lexicales « nouvellement déplacées », nous avons choisi de nous intéresser uniquement à celles portant sur les noms, les verbes, les adverbes et les adjectifs. Les pronoms, les déterminants et les conjonctions ont été écartés puisque nous avons pensé intuitivement, entre autres, qu'ils se prêteraient moins aux types de rapports que nous cherchions à identifier. Cela étant, nous avons établi deux distinctions. La première distinction renvoie aux erreurs lexicales qui s'avèrent inacceptables au vu des « rapports de force » entre les items lexicaux à un niveau intra-phrastique. C'est-à-dire qu'un autre terme entretenant une relation collocationnelle avec ceux disponibles dans la phrase aurait mieux convenu à la situation, rendant le choix effectif de l'apprenant quelque peu inopiné voire biscornu. Au-delà des rapports collocationnels qui auraient pu exister, la deuxième catégorie signifie tout simplement que l'item conduit de manière générale à un non-sens – indépendamment de la grammaticalité de la phrase.

[127] La façon dont nous avons obtenu ce chiffre est expliquée en section 8.3.1.2.

Etant donné que les erreurs lexicales portant sur les verbes ont été les plus nombreuses, nous avons choisi ces quatre exemples suivants pour illustrer nos propos. Les deux premiers renvoient au premier cas de figure, à savoir que le terme est inacceptable vis-à-vis du rapport entretenu avec son complément. Les deux derniers sont par conséquent des erreurs de sens, stricto sensu.

104. All countries were **touched* $affected$ by that crisis […]. (txt_015_sm)
105. The situation **emphasized* $##$ the problem of global warming and doesn't **enhance* $##$ the risk of sea level rising. (txt_027_txt_sm1)
106. But, on the other hand, I am **convicted* $convinced$ that the economy would find other ways to bridge this gap. (txt033_sm1)
107. But the *dept[sic]* relief of these countries can **wonder* $##$ several years and the reduction of public spending [...] (txt_048_sm1)

	rapport	sens	*total (%)*
nom	70	30	*100*
verbe	68,05	31,95	*100*
adverbe	50	50	*100*
adjectif	96,6	3,4	*100*

Tableau 50 : Erreurs lexicales simple et phraséologique

Le tableau 50 résume les résultats de nos analyses complémentaires. En effet, on note parmi les erreurs lexicales dans le tableau qu'un nombre considérable semble être le fruit d'un rapport de force préexistant entre les mots dont les apprenants n'auraient pas encore connaissance. A titre d'exemple, 96,6% des adjectifs signalés comme erronés ne relèvent pas simplement d'une erreur sémantique banale mais proviennent du fait que l'adjectif n'est pas prédisposé à jouer le rôle de modifieur pour lequel il est employé. Le pourcentage est également notable pour le verbe et pour le nom, à savoir qu'autour de 70% relève du rapport de force méconnu entre les termes employés. Ces observations nous permettant déjà en l'état de repenser l'appellation des occurrences – situées dans la colonne intitulée « rapport » – désormais comme des erreurs de phraséologie lexicale et non seulement en tant qu'erreurs lexicales simples.

A ce titre, il y aura un rapprochement indéniable à faire entre les erreurs individuelles identifiées ici et celles qui sont issues d'unités multi-mots (ou complexes) que l'on a pu observer précédemment dans les sections 6.1.2 et 6.2.1. Soulignons en effet que l'ensemble de ces erreurs traduisent une sorte de défigement à la fois des expressions figées ou semi-figées en langue anglaise, mais également un défigement entre des éléments qui ne relèvent pas des expressions connues en tant que telles de la langue mais qui dans l'usage fonctionnent de manière similaire. Le fait de savoir que les erreurs lexicales ne sont pas aléatoires, comme nous l'avions initialement pensé, et qu'elles

renvoient à un niveau de difficulté supplémentaire auquel les apprenants doivent faire face mérite que l'on y accorde une attention particulière en cours de langue. Mais au vu de la singularité des rapports proprement phraséologiques et des différents types connus (cf. Cowie 1998 ; Granger & Paquot 2008 ; Osborne 2008 ; Thewissen 2008), la manière d'aborder ce problème de façon concrète reste à définir.

8.2.2.2 Les erreurs de parataxe et les structures asyndétiques

Le deuxième écueil que l'on va traiter ici renvoie à un phénomène que peu de didacticiens abordent comme relevant d'un obstacle potentiel dans la production écrite en anglais langue étrangère. Il s'agit de la construction dite paratactique et plus singulièrement de la structure asyndétique. En effet, dès lors que l'on rédige un texte, on doit « jongler » de façon plus au moins inconsciente entre des constructions hypotactiques et des constructions paratactiques. La première permet de hiérarchiser les informations ou plus précisément facilite la subordination des propositions à un niveau inter- et intra-phrastique, tandis que la deuxième permet la coordination d'éléments appartenant en principe au même rang. Et c'est justement sur cette dernière que l'on relève un certain nombre de difficultés chez nos sujets-participants.

Au fur et à mesure que nous annotions le corpus, nous avons constaté que quelques apprenants rencontraient un problème dans la coordination des éléments de la phrase. Et une fois l'étape de l'annotation terminée, nous nous sommes intéressé à toutes les catégories susceptibles de nous apporter des éclaircissements sur ce point. A ce titre, nous nous sommes tourné vers l'ensemble des items étiquetés en tant qu'erreurs de conjonction, de connecteur voire de ponctuation, dans le but de pouvoir identifier la source du problème. Après un examen minutieux des cinq catégories indiquées ci-dessous, la structure asyndétique a été reconnue comme étant un véritable écueil pour un certain nombre de nos apprenants.

	sm1	**sm2**	**n**
incorrect-clause-conjunction	23	21	44
missing-clause-conjunction	46	31	77
punctuation-required-not-present	20	14	34
punctuation-inserted-not-required	26	7	33
missing-connector	34	29	63
	149	102	251

Tableau 51 : Les catégories utilisées pour étudier la parataxe

L'ensemble des 251 étiquetages a été passé en revue. La première catégorie dans le tableau 51, à savoir l'étiquetage portant sur le choix de conjonction inapproprié, ne s'est pas révélée concluante.

Par contre, les quatre autres nous ont permis de faire deux observations. Tout d'abord, l'absence de conjonction que l'on a pu observer recouvre des caractéristiques similaires d'un apprenant à un autre. Ce dernier cherche à coordonner des éléments qui ne sont pas d'un rang identique ou ne renvoient pas au même type d'information, voire ne sont tout simplement pas adaptés à ce type de coordination. 81,8% des items repérés sont similaires à l'exemple ci-dessous.

108. To conclude the carbon offset is obviously better than nothing, but people could do better for the environment *an[sic]* the prevention of global warming increase; **higher standards, renewable energy, sustainable development, real reduction of greenhouse gases emissions* $##$. (txt_003_sm1)

La deuxième observation concerne les trois autres catégories du tableau qui n'ont pas apporté de pourcentages aussi probants que celle que l'on vient d'examiner. Mais bien qu'elles ne soient pas toutes clairement asyndétiques, notons tout de même qu'elles recouvrent des erreurs de coordination et de subordination à cheval entre la parataxe et l'hypotaxe. En effet, le fait que la relation ne soit pas clairement établie fait écho aux erreurs d'ostension que nous avons signalées dans la chapitre VI, comme relevant de l'ordre de l'acceptabilité textuelle. Pour rappel, ces erreurs surviennent quand l'apprenant n'a pas apporté suffisamment d'informations nécessaires pour que le lien entre ses propos et l'exemple qu'il a fourni soit formellement établi. De plus, cette erreur prend souvent la forme d'un exemple détaché et donc isolé qui est placé directement après une proposition. La responsabilité appartient à celui qui lit le texte de faire le lien.

109. Women are the primary *breadwinner[sic]* in the house: even if they have nannies, it's not *enough efficient[sic]* to help them to reach self-fulfilment. Whereas men would be outward-looking: **quest of power* $##$. (txt_024_sm2)

En définitive, les erreurs observées à ce niveau sont soit de l'ordre du premier exemple, soit du deuxième. Et on est vraisemblablement face à un problème qui nécessiterait que l'on y accorde une attention particulière en cours de langue. Reconnaissons toutefois ici que ce point mériterait une étude complémentaire. Peut-être gagnerait-on davantage à s'intéresser à un corpus d'apprenants dans le but précis d'étudier les structures hypotactiques et paratactiques, et ce, de façon à approfondir l'analyse issue des catégorisations que nous avons établies.

8.2.3 Quelques réflexions sur les résultats

En ce qui concerne l'ensemble des erreurs que nous avons appelé « acceptabilité textuelle » (cf. chapitre VI) ou « calcul sémantique » (cf. section 8.2.1), nous sommes conscient que certains verront dans ces éléments le résultat pur et simple d'une non-maîtrise grammaticale. C'est-à-dire,

nous pensons qu'un des arguments que l'on pourrait avancer contre ces catégories est que ces erreurs disparaitront avec une maîtrise grammaticale solide. C'est pourquoi nous soulignons que la suite inévitable de notre étude est de poursuivre l'analyse avec des profils d'apprenants avancés confirmés – de manière à vérifier si les erreurs du système linguistique et les erreurs d'acceptabilité textuelle persisteront dans un tel groupe avec des niveaux homogènes. Par contre, nous nous attendrons à ce que le nombre d'erreurs comptabilisées soit significativement différent de ce que nous avons obtenu dans la présente étude – en raison notamment de l'hétérogénéité des profils et du nombre de nos sujets-participants qui se sont avérés relever d'un faible niveau en langue anglaise – avoisinant le niveau B1.

A cette première remarque sur les erreurs d'acceptabilité textuelle et de calcul sémantique s'ajoute celle portant sur les constructions paratactiques et hypotactiques (cf. section 8.2.2.2). Bien que nous n'ayons pas avancé d'hypothèse spécifique sur la cause de ces erreurs, nous reconnaissons que plusieurs explications sont possibles. Par exemple, les erreurs peuvent relever (i) d'un transfert (négatif) de style de la langue française vers la langue anglaise ; (ii) d'une compétence langagière générale non-maîtrisée en L1 qui se voit transférée en l'état vers la L2 ; voire (iii) d'une erreur propre à l'interlangue en cours de développement. Etant donné la multiplicité d'options possibles, il nous parait judicieux de ne pas soumettre d'hypothèse prématurée et d'attendre d'avoir une étude ultérieure dans laquelle ces structures constitueraient l'objet central d'analyse.

En somme, la multiplicité des catégories d'erreurs textuelles de la section 6.1 et des erreurs d'acceptabilité textuelle de la section 6.2 témoignent du fait que l'analyse des erreurs ne doit plus se contenter de faire de l'aspect grammatical le seul point d'arrivée et de départ de son cadre analytique. Autrement dit, la correction grammaticale ne doit pas être le seul objectif de l'analyse – de la même manière qu'elle ne peut pas être le seul objectif d'un cours de langue étrangère. En effet, malgré les chevauchements que nous avons mis en avant dans la section 6.1.4, nos résultats ont permis de démontrer que certaines erreurs ne peuvent pas être élucidées si l'on se cantonne à la grammaire traditionnelle : à savoir, par exemple, les erreurs référentielles, les différents types d'erreurs de focus (notamment, l'ostension et la progression thématique), sans oublier bien entendu, les erreurs de mise en phrase et de cadrage. Dans cette optique, ne pas prendre en compte le niveau textuel dans l'analyse des erreurs signifie d'emblée écarter un certain nombre de phénomènes linguistiques intéressants qui ne pourraient être expliqué qu'à un niveau proprement textuel (cf. à ce titre, notre schéma explicatif dans la section 8.3.2, dans lequel on distingue formellement entre les erreurs qui peuvent être expliquées selon le système linguistique et le texte).

Et par voie de conséquence, nous soutenons que cette mise à l'écart aura pour résultat de fausser tout simplement l'analyse, ou du moins fournir à l'analyste des données parcellaires.

8.3 Des comparaisons avec d'autres études et notre modèle de restructuration

Passons maintenant à la comparaison de nos résultats avec ceux issus de quatre études différentes. Tout d'abord, l'interface entre ce que l'on considère comme étant des erreurs lexicales, syntaxiques et sémantiques sera exploitée (cf. 8.3.1) ; ensuite, nous présenterons un cadre explicatif de l'ensemble des erreurs que nous avons évoquées dans la présente étude.

8.3.1 L'interface entre erreurs lexicales, syntaxiques et sémantiques

En explorant la littérature sur l'analyse des erreurs, on ne peut s'empêcher de remarquer l'abondance des différences conceptuelles attribuées au terme « erreur », dont certaines ont été examinées dans les sections 1.3.1 et 1.3.2. Cette profusion de définitions ne se limite pas au terme lui-même – elle est également présente à travers les catégorisations et les descriptions linguistiques, sans oublier les étiquetages dans l'annotation informatique, où chaque cadre conceptuel (ou chaque projet d'analyse) semble donner lieu à des nouvelles terminologies. Et bien que cette tendance traduise dans un premier temps une sorte de dynamisme du champ de l'analyse des erreurs, il devient difficile d'effectuer des analyses comparatives lorsque la terminologie est non seulement différente, mais aussi quand les sous-catégories ne sont pas souvent directement transposables les unes aux autres. Pour cette raison, nous allons nous pencher sur les études qui ont des taxonomies d'erreur semblables aux nôtres et/ou des paramètres de recherche similaires en termes de profils d'apprenants. A ce titre, nous allons nous référer aux trois types d'erreur qui, à notre sens, sont les plus communément reprises d'une étude à l'autre : à savoir des erreurs lexicales, grammaticales (au sens syntaxique) et sémantiques.

8.3.1.1 Les erreurs lexicales

La première remarque qui s'impose ici concerne le schéma d'annotation d'erreurs intégré au logiciel d'UAM CorpusTool. Ce schéma que nous avons employé dans le premier volet de notre annotation, rappelons-le, fournit une taxonomie très développée sur les différentes catégories d'erreurs les plus souvent exploitées dans la littérature. Or, cette « recherche de l'exhaustivité », qui à première vue est appréciable, s'avère considérablement contraignante. En effet, du fait des différents niveaux de profondeur et de l'aspect très pointilleux des schémas d'annotation

correspondants, nous jugeons leur comparaison avec d'autres projets ayant des taxonomies plus restreintes plutôt problématique.

Soit il faut « remonter » le niveau de profondeur pour effectuer des comparaisons de manière « à peu près convenable » ; soit il faut vraisemblablement « mutiler » l'arborescence et sélectionner des étiquetages individuels. Mais notons que même cette approche n'est pas satisfaisante puisque de temps en temps l'étiquetage qui se trouve en fin d'un nœud donné (au sens d'une branche sur l'arbre taxonomique) ne partage pas la même base ou racine dans un autre schéma – ce qui rend la comparaison encore plus délicate. Cette observation se veut pour l'ensemble des études que nous allons comparer dans cette section et les deux autres sections qui vont suivre (c'est-à-dire dans les sous-sections 8.3.1.2 et 8.3.1.3). Mais elle est tout particulièrement le cas pour le premier projet que nous décrivons ci-après.

Perry (1993)

Une étude qui a attiré notre attention en raison des nombreux paramètres contextuels similaires à la nôtre mais qui s'est ensuite révélée incomparable est celle de Perry (1993). En effet, parmi les paramètres partagés, notons que le contexte universitaire de la collecte du corpus est identique, de même que les types de textes recueillis et les profils des apprenants : à savoir, entre autres, des étudiants en première année à l'université Paris Dauphine. Mais contrairement à notre projet, l'auteur a effectué une analyse des erreurs sur 44 textes, en accordant une attention particulière à ce qu'elle considère comme des erreurs lexicales.

L'auteur a ensuite procédé à un classement des six catégories principales – qui, à notre sens, est intuitivement représentatif de ce que nous avons vu à la fois en tant qu'enseignant dans le même établissement mais également à travers ce que nous avons observé dans notre corpus. Ce qui pose vraisemblablement problème pour la comparaison est que les catégories lexicales que Perry a mises en avant ne sont pas clairement définies : à savoir (i) confusion entre L1/L2 ; (ii) confusion phonologique en L2 ; (iii) choix lexical inapproprié ; (iv) collocation L2 ; (v) collocation L1 et enfin (vi) erreurs lexico-grammaticales.

Si on les examine de près, aucune de ces catégories ne correspond à nos étiquetages ni dans la terminologie ni dans le sens visé. De plus, les catégories n'ont pas été suffisamment définies de façon à pouvoir saisir les différences subtiles qu'il pourrait y avoir entre ces six termes. A titre d'exemple, nous avons annoté un nombre considérable d'erreurs d'orthographe qui pourraient à la fois être classées dans le premier ou deuxième groupe de Perry. A ceci s'ajoute aussi diverses

erreurs annotées en tant qu'emprunts ou « transferts/calques » qui pourraient tout aussi bien entrer dans plusieurs de ces groupes. Cela étant, sans une explicitation formelle de ces catégories, toute comparaison nous semble hasardeuse.

Hemchua & Schmitt (2006)

Dans l'étude de Hemchua & Schmitt (2006) les apprenants sont de langue maternelle thaïe (thaïlandaise) et sont inscrits en licence d'anglais à l'université. Comme dans l'étude de Perry, les auteurs se sont exclusivement intéressés aux erreurs lexicales. Mais cette fois-ci, les catégorisations relèvent de celles qui sont le plus communément faites[128] : à savoir d'une part les erreurs lexicales formelles qui renvoient principalement à des notions de « misselection », « misformation », « distortion »[129]; et les erreurs lexicales dites purement sémantiques d'autre part. Parmi les 24 étiquetages possibles, nous n'en avons que quatre en commun qui soient directement transposables et quatre de plus qui pourraient correspondre « plus au moins » à une catégorie supplémentaire de notre schéma d'annotation. Ces éléments sont comparés ci-dessous.

	Hemchua & Schmitt (%)	Notre étude (%)
borrowing	0,00	1,35
coinage	0,00	2,01
calque	6,90	8,41
false-friend	1,15	0,91
semantic (1.1 – 1.4)[130]	24,9	38,1

Tableau 52 : Comparaison des erreurs lexicales entre deux études

Dans le tableau 52 les chiffres indiquent le pourcentage de l'item par rapport à l'ensemble des erreurs lexicales observées. Ces comparaisons permettent d'établir un certain parallèle dans les catégories examinées et appellent deux remarques. La première est que les écarts sont a priori minimes mises à part deux catégories qui n'ont enregistré aucune erreur, ce qui pourrait s'expliquer par la différence entre la langue maternelle des apprenants. En effet, le français est sensiblement plus proche de l'anglais ce qui pourrait favoriser les erreurs des deux premiers groupes (des transferts donc du français vers l'anglais), tandis que la langue thaïe est plus éloignée de la langue cible pour ce qui est des participants de Hemchua & Schmitt. On pourrait alors croire qu'il y aurait moins de tentative ou de possibilité de ce type de transfert.

[128] Cf. James (1998) qui retrace les différentes catégorisations d'erreurs lexicales.
[129] Ces trois termes seront expliqués en détail dans la section 8.3.3.
[130] Les erreurs sémantiques lexicales numérotées *1 .1* à *1.4* par Hemchua & Schmitt (2006 : 15) sont comparées à titre illustratif à nos différentes erreurs de vocabulaire préalablement détaillées dans le chapitre V.

Nous pensons cependant que l'écart observé peut également s'expliquer de façon plus pratique si l'on considère que les auteurs n'ont étudié que 20 textes et n'ont par conséquent obtenu que 261 occurrences annotées. Ce faible effectif général permet donc de relativiser les « 0,00 » occurrences observées, tout en soulignant davantage la similitude manifeste des résultats. En bref, il semblerait que les erreurs de transfert de sens soient beaucoup moins fréquentes que les erreurs de vocabulaire[131] qui constituent, à notre sens, un problème particulièrement épineux pour les deux groupes d'apprenants comparés. De plus, ces résultats nous laissent penser qu'à un niveau de maîtrise égal, la typologie des erreurs lexicales sera similaire, indépendamment de la langue maternelle des apprenants.

On pourrait ajouter ici plusieurs autres études portant sur les erreurs lexicales, mais le constat sera indéniablement le même. Les catégories sont souvent tellement éloignées les unes des autres et, de ce fait, ne se prêtent pas à des comparaisons optimales. Ce constat est « fort dommage » pour des projets ayant un certain nombre de paramètres conceptuels identiques, puisque la comparaison pourrait permettre d'un point de vue didactique de mieux rendre compte de certains phénomènes d'erreurs observés entre des groupes différents. Par exemple, si la même taxonomie est utilisée pour étudier plusieurs groupes de participants apprenant la même langue étrangère mais ayant des langues maternelles différentes, cela pourrait faciliter une meilleure compréhension des enjeux de l'acquisition en indiquant ce qui relève d'un point problématique commun à plusieurs apprenants et ce qui est propre à d'autres.

8.3.1.2 Les erreurs morphosyntaxiques

Une autre catégorie qui semble avoir des frontières conceptuelles ambiguës est celle des erreurs morphosyntaxiques. Ces erreurs varient considérablement en effet selon la terminologie utilisée pour délimiter leur utilisation et les nombreux étiquetages qui se veulent « tous inclusifs » mais qui ne constituent, à notre sens, qu'une sorte de fourre-tout. Dans cette perspective, le schéma d'annotation d'UAM CorpusTool range certaines erreurs tout d'abord en fonction de leurs classes ou catégories grammaticales - les sous-types ou sous-groupes sont ensuite classés au vu d'une catégorisation approfondie issue d'une granularité fine (fined-grained). Ce que nous avons remarqué en plus, est que les balises sous « erreurs grammaticales » comptent pour 75% de l'ensemble des six types d'erreurs disponibles dans le schéma : à savoir plus de 90 étiquetages

[131] La notion d'erreur de vocabulaire est souvent interchangeable avec erreur lexicale et erreur sémantique (cf notamment les sections 5.1.1.6, 6.2.1 et 8.2.2) Nous allons donc essayer d'établir une distinction formelle entre l'ensemble de ces erreurs dans la section 8.3.2.

uniquement pour cette catégorie. Cependant, comme nous l'avons expliqué dans la section précédente (cf. section 8.3.1.1) certains des sous-types d'erreurs grammaticales auraient gagné à être rattachés ailleurs, par exemple dans la catégorie d'erreur lexicale.

Par ailleurs, de nombreuses autres études ont tendance à éviter des termes comme « erreurs morphosyntaxiques » qui regroupent le plus grand nombre d'erreurs comptabilisées à ce niveau, en optant souvent pour des équivalents un peu plus larges ou plus vagues comme par exemple « erreurs grammaticales ». Toutefois, certaines tentent de faire l'inverse en poussant la différenciation « à l'extrême » à l'instar de Granger & Monfort (1994), Dagneaux et al. (1998) et Chuang & Nesi (2006) qui font la distinction, à titre illustratif, entre des erreurs lexicales, des erreurs grammaticales et des erreurs lexico-grammaticales. Le résultat de cette profusion terminologique est résolument le même que dans la section précédente. C'est-à-dire qu'il devient de plus en plus difficile de réaliser de manière systématique et exhaustive des analyses comparatives avec d'autres études : l'analyste a désormais la responsabilité d'identifier de manière objective un très faible nombre de catégories parmi une large sélection qui sont définies de manière approximativement similaire dans au moins deux projets utilisant souvent des terminologies diamétralement opposées. De plus, la granularité des étiquetages n'aurait pas forcément la même profondeur ou le même séquençage. La tâche de l'analyste comporte, de ce fait, une part considérable de subjectivité dans l'interprétation et dans les rapprochements faits entre les différents étiquetages.

Ce constat semble a priori partagé par Dagneaux et al. qui nous rapportent qu'en plus d'avoir des catégories dont les contours ne se distinguent pas facilement, des étiquetages très « pointus » pourraient également conduire à avoir plusieurs possibilités d'annotation sur un même item.

> Descriptive categories such as these are not enough to ensure consistency of analysis. Researchers need to know exactly what is meant by 'grammatical' or 'lexico-grammatical', for instance. In addition, they need to be provided with clear guidelines in case errors […] allow for more than one analysis. (1998: 166)

Cela suppose que l'analyste soit obligé de lire le manuel d'annotation afin de saisir les subtilités entre les termes. Et bien que l'on comprenne le raisonnement et la démarche liés à ce besoin d'avoir une certaine cohérence dans les annotations et par extension dans les études comparatives, avoir des étiquetages avec des contours conceptuels flous ne facilite aucunement la comparaison. Il y a même un risque, au contraire, d'éloigner ceux qui veulent éviter ou écarter des biais d'interprétation s'ils venaient à comparer l'incomparable entre deux études. Cela étant dit, nous

allons tout de même tenter une comparaison avec quelques éléments de l'étude de Dagneaux et al. (1998), en raison notamment du fait qu'elle partage bien des traits caractéristiques avec la nôtre : à savoir, entre autres, que leurs participants sont de langue maternelle française comme nos sujets-participants.

	ICLE [FR]	UAM [1er]	UAM [2ème]
lexical	30%	18,6%	34,8%
lexico-grammatical	7%	57,4%	41,1%
grammatical	32%		

Tableau 53 : Comparaison entre ICLE et notre étude (lexique et grammaire)

Si l'on compare trois des sept (macro)catégories proposées dans Dagneaux et al. aux deux catégories qui s'en rapprochent le plus dans notre étude, il devient alors évident que malgré les dissimilitudes que l'on a décrites ci-dessus, on est vraisemblablement face à des phénomènes similaires. Par exemple, en comparant les erreurs lexicales, lexico-grammaticales et grammaticales d'un des corpus d'ICLE (FR) avec celles mises en avant dans le chapitre V (UAM [1er]), le tout semble à première vue assez éloigné l'un de l'autre. Mais en prenant en compte les considérations terminologiques derrières la notion d'erreur lexicale chez Granger & Monfort (1994), on arrive à déplacer 16% des erreurs - préalablement rangées dans la catégorie grammaticale dans le schéma d'UAM – vers la catégorie lexicale, ce qui d'emblée permet d'observer une tendance qui se rapproche entre ICLE [FR] et UAM [2ème].

	ICLE [FR]	UAM [1er]	UAM [2ème]
déterminant	27%	20%	27%
verbe	24%	12,7%	17,7%
pronom	24%	7%	9,8%
total[132]	*75%*	*39,7%*	*54,5%*

Tableau 54: Comparaison entre ICLE et notre étude (les erreurs les plus fréquentes)

Si l'on pousse la comparaison un peu plus loin, on pourrait également s'intéresser aux trois points les plus problématiques pour les apprenants dans l'étude de Dagneaux et al. Le tableau ci-dessus montre la répartition des erreurs par rapport à l'ensemble de ce qui a été annoté dans la catégorie grammaticale. UAM [1er] et UAM [2ème] correspondent respectivement aux résultats initiaux tels que mis en avant dans le chapitre V (avec un total de 3107 erreurs grammaticales) et aux résultats après avoir enlevé de cette catégorie, ce que Dagneaux et al. considèrent comme étant purement lexical (moins 880 occurrences). Entre ICLE [FR] et UAM [1er], les erreurs de déterminant sont ce

[132] Rappelons qu'uniquement quelques exemples ont été choisis, ce qui fait que le total ne fait pas 100%.

qu'il y a de plus récurrent entre les deux études. Entre ICLE [FR] et UAM [2ème] il y a deux catégories identiques, ce qui signifie à notre sens que deux premiers types d'erreurs constituent des véritables écueils pour les apprenants francophones. Bien que le pourcentage diffère énormément pour ce qui est des erreurs portant sur le pronom, il nous semble judicieux de rappeler ici que le schéma d'annotation d'UAM comporte plus de 90 étiquetages dans la catégorie grammaticale contre moins de 60 pour l'ensemble du schéma d'ICLE, ce qui pourrait en partie expliquer les écarts de pourcentages observés notamment sur les erreurs de pronom. A cette remarque s'ajoute le fait que certaines erreurs de pronom ont été signalées séparément dans notre étude comme des erreurs de référence (cf. section 6.1.1), et cette séparation n'existe pas dans l'annotation ICLE.

Toutefois, il faut souligner qu'il y a de nombreux étiquetages pour lesquels la comparaison est tout simplement impossible en raison du fait que les balises sont souvent « trop précises » ou au contraire ont des contours conceptuels bien trop vagues. Nous ne pouvons que regretter de ne pas pouvoir mener une comparaison plus exhaustive, sans risquer de tomber dans des approximations hasardeuses. En effet, une véritable comparaison point par point aurait permis d'avoir une vue plus globale des erreurs qui posent problème aux deux groupes de participants francophones et cela aurait également facilité la généralisation des résultats des deux études selon les différents points de convergence. Hélas, nous ne pouvons sélectionner que quelques éléments pour effectuer une réelle comparaison, les résultats restants ne demeurent valables que pour l'étude dans laquelle ils ont été observés.

8.3.1.3 Les erreurs sémantiques

Le débat sur les erreurs sémantiques est sensiblement le même que dans les deux sous-sections précédentes. Certains (cf. Cabassut 2003) utilisent le terme d'erreur sémantique ou d'erreur lexicale pour un même phénomène et donc de façon interchangeable tandis que d'autres (cf. Laufer 1994 ; Granger & Monfort 1994 ; Anctil 2010) voient dans ces deux appellations des faits très distincts. Force est de constater cependant que peu d'études parmi celles que nous avons consultées ont employé le terme « erreur sémantique » et font, de ce fait, un amalgame entre les deux types d'erreurs cités ci-dessus. Pour ce qui est de notre ressort, il semble important de bien les distinguer puisque la source et les explications qui en résultent sont totalement différentes selon l'approche que l'on adopte. En effet, l'erreur sémantique renvoie à l'axe paradigmatique et ne doit prendre en compte que les erreurs dont le seul problème vient du sens final obtenu en contexte : l'erreur sémantique serait, de ce fait, permutable avec un autre item appartenant à la même catégorie grammaticale. Nous reviendrons à la différenciation de ces deux types d'erreurs dans la section

suivante, et nous proposerons une structure hiérarchisée permettant de bien catégoriser l'ensemble des erreurs que nous avons examinées jusqu'ici.

8.3.2 Vers une restructuration [de la prise en charge] des erreurs

Après avoir examiné les différentes erreurs relevées dans notre corpus et les avoir comparées à d'autres projets d'analyse d'erreurs, notre constat final est le suivant. Des dissimilitudes ont été signalées à plusieurs niveaux, notamment dans les nombreuses taxonomies que nous avons pu consulter. Toutefois un certain nombre de ces projets se rejoignent au niveau de l'interprétation et de l'explication finale des phénomènes observés. Et bien que nous nous soyons rangés derrière les explications souvent très détaillées, il nous semble que ces explications manquent de systématicité dans leur approche tout aussi bien que d'un cadre permettant d'avoir un aperçu global des différents types d'erreurs observés et observables – le tout aligné selon la typologie mais également l'explication de ce qui les provoque. A cet effet, nous proposons le cadre ci-après (cf. figure 40), qui se veut hiérarchisé et extensible.

Indépendamment des choix taxonomiques, des niveaux de profondeur accordés aux étiquetages sur un schéma d'annotation ou de la classe grammaticale à laquelle appartient une erreur donnée, nous estimons que tout signalement d'erreur trouve un début d'explication dans cette hiérarchisation. Au-delà du fait que notre modèle se veut « inclusif », il permet de rendre compte des phénomènes aussi bien individuels que multiples. Plus précisément, il prend en compte les erreurs portant sur des unités lexicales individuelles [ULI] et des unités lexicales multiples ou multi-mots [ULM]. Cet aspect découle principalement du besoin identifié dans le présent travail d'avoir un moyen de porter une attention particulière sur les erreurs situées à plusieurs niveaux du discours ou de texte, et ce de manière simultanée. De ce fait, le cadre facilite la description des plus petites unités des sens qui seront identifiées en tant qu'erreurs jusqu'aux empans textuels plus larges.

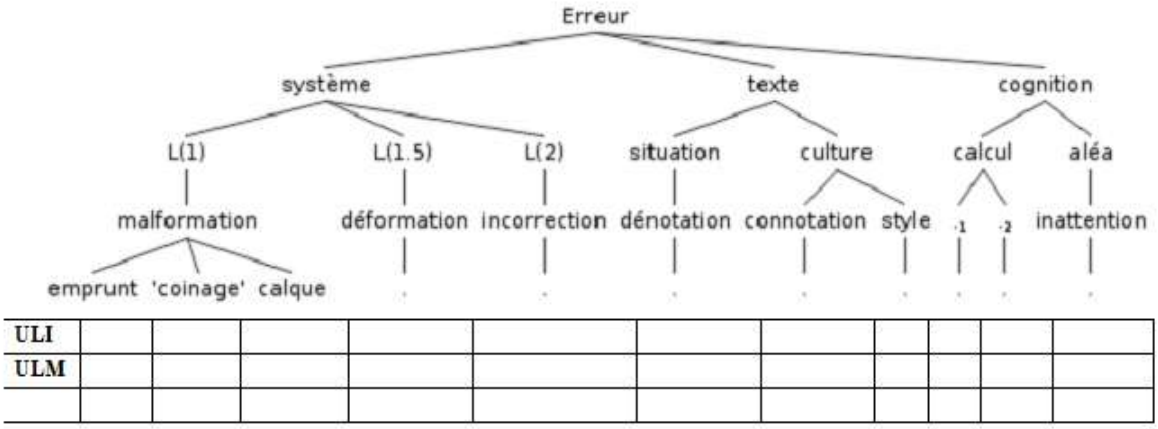


Figure 40 : proposition de schéma explicatif des erreurs

Passons maintenant à la description des éléments à l'intérieur du cadre. L'étude des erreurs nous a conduits à faire un premier constat qui se veut global. Hormis les erreurs imputables aux aléas accidentels de la vie, une erreur en langue étrangère peut trouver une explication (i) au niveau des systèmes linguistiques connus de l'apprenant et qui entretiennent des rapports de force les uns avec les autres ; (ii) au niveau du contexte général de l'emploi desdits systèmes ; et (iii) au niveau de la demande cognitive, en termes de calcul sémantique. Les systèmes ici sont compris comme étant la langue maternelle [L(1)], la langue cible [L(2)], et la langue en cours d'apprentissage ou de développement [L(1.5)]. Notons que cette dernière peut revêtir des traits propres à la [L(1)] et la [L(2], mais peut tout aussi bien avoir des caractéristiques propres, qui ne sont identifiables dans aucun des deux autres systèmes mentionnés. Soulignons ensuite que le texte (y compris le contexte textuel lui-même) est tout particulièrement important dans cette description (i) puisqu'il permet d'expliquer les jugements d'acceptabilité d'un item qui pourrait en temps normal être tenu pour acceptable dans la [L(1)] ou la [L(2] et (ii) puisqu'il permet également d'expliquer ce qui pose vraisemblablement problème au vu d'un emploi (con)textuel erroné. Enfin, l'aspect cognitif ne doit pas être écarté – dans la mesure où certaines erreurs ne sont ni totalement imputables aux systèmes linguistiques, ni totalement imputables aux exigences textuelles, mais relèvent d'un chevauchement d'ordre cognitif.

Les termes malformation, déformation et incorrection dans la figure 1 renvoient respectivement, dans une certaine mesure, à ce que d'autres appellent « misformation », « distortion », « misselection » (cf. James 1998 ; Chuang & Nesi 2006 ; Hemchua & Schmitt 2006). Toutefois, ces termes n'ont ni été mis en rapport tels quels avec le système linguistique (c'est-à-dire, L(1), L(1.5), ou L(2)) qui régit leur apparition ni été utilisés pour expliquer autre chose que des erreurs lexicales individuelles. Pour ce qui est de la malformation, nous reprenons les trois sous-types de James (1998) : il s'agit d'abord d'une erreur « d'emprunt » qui est transféré de la langue source vers la

langue cible sans modification ; « coinage » renvoie à un emprunt qui se voit adapté à la morphosyntaxe de la langue cible ; enfin « calque » est une transposition littérale d'un terme d'une autre langue vers la [L(2)]. Pour rappel, l'ensemble de ces descriptions s'applique dans le cadre présenté à la fois aux erreurs portant sur des unités lexicales individuelles et sur les unités lexicales multiples ou multi-mots. Quelques exemples sont fournis pour illustrer ces trois types d'erreurs.

emprunt	ULI	[…] without including any **changement* $changes$ in […] (txt_006_sm2)
coinage	ULI	[…] *to *applicate* $apply$ their skills in an exam […] (txt_023_sm2)
	ULM	[...] this action can **in contrary* $on the contrary$ decrease consumers' consumption and [...] (txt_026_sm1)
calque	ULI	[…] able to study some **formations* $courses$ […] (txt_022_sm2)
	ULM	[…] and even before the **superior studies* $university studies$. (txt_023_sm2)

Tableau 55 : Quelques exemples d'erreurs de malformation

Pour ce qui est de la déformation, les items erronés n'existent en tant que tels ni dans la langue maternelle ni dans la langue cible. Cependant, il est souvent possible de faire un rapprochement avec une des deux.

déformation	ULI	[…] more than 80% of girls take **litterare* $literature$ courses […] (txt_028_sm2)
	ULM	Even if they are **high-graduated* $###$ […] (txt_024_sm2)

Tableau 56 : Quelques exemples d'erreurs de déformation

L'incorrection renvoie à quelque chose d'existant dans la langue cible mais qui, à défaut de conduire à une erreur grammaticale, conduit à un non-sens. Il s'agit donc de ce que l'on pourrait appeler une erreur lexicale.

incorrection	ULI	[…] boy [...]have to show **insurance* $assurance$ when facing facing female gender […] (txt_028_sm2)
	ULM	[…] compelled to **make new studies* $carry out new studies/research$ in another field. (txt_113_sm1)

Tableau 57 : Quelques exemples d'erreurs d'incorrection

Notons qu'il y a un léger chevauchement entre incorrection et dénotation. Cette dernière renvoie également à un choix possible dans la langue cible, mais ne peut être jugée sémantiquement inappropriée qu'au vu du contexte textuel dans lequel elle se trouve. Les véritables erreurs lexico-sémantiques ont, de ce fait, toute leur place dans cette catégorie.

dénotation	ULI	The lack of agreement […] only **emphasizes* $calls attention to$ the issue.

		(txt_004_sm2)
	ULM	But it is really the case in *_the deep down_ $##$? (txt_084_sm1)

Tableau 58 : Quelques exemples d'erreurs de dénotation

Quant aux erreurs de connotation et de style, le système linguistique n'est pas à proprement parler mis en cause. C'est tout singulièrement le sens qui est inapproprié dans connotation par rapport, par exemple, au genre textuel. Il peut par conséquent être question ici d'une erreur de registre.

connotation	ULI	*_Cause_ $Because$ on the other side of the coin, countries and companies […]. (txt_022_sm1)
	ULM	There are *_loads of_ $several$ issues to address […] (txt_004_sm2)

Tableau 59 : Quelques exemples d'erreurs de connotation

Le style en termes de choix personnel peut tout aussi bien être la cause d'un jugement d'acceptabilité. Dans ce cas, il est imputable au contexte culturel de son emploi. A titre illustratif, comme nous l'avons vu dans le chapitre II, certaines cultures discursives privilégient des textes avec un style orienté vers le lecteur – ce style est, de ce fait, considéré plus « pédagogue » –, tandis que d'autres apprécient davantage la « richesse informationnelle », sans accorder une importance considérable à la forme ou la macrostructure adoptée par le scripteur. Des jugements de valeur peuvent donc porter sur le style à un niveau proprement textuel ou à un niveau proprement intra-phrastique.

Enfin, les erreurs inter(-1) et intra texte (-2) renvoient à des calculs sémantiques. Pour la première, elle survient lors que l'on constate que la somme des propriétés syntaxiques d'un énoncé (ou d'un segment phrastique) n'est pas transférée sur l'élément (d'un autre énoncé) qui – en temps normal – aurait dû porter la marque de ces « calculs ». Pour la deuxième, le report de la somme est constaté à un niveau intra-phrastique. D'une manière générale, les items que l'on relève ici renvoient, entre autres, aux différents types d'erreurs de référents et aux erreurs de concordance de temps.

calcul	inter	Besides women are expected to do the majority of *this[sic]* household tasks, preparing the meal, bringing up children ... So it seems to be difficult for *_her_ $them$ to balance work and family. (txt_056_sm2)
	intra	-[…] half of the candidates of each election in each political organisation must be women. It seems to be a main breakthrough, but *_if only[sic]_* it *_has_* worked $only if it works$! (txt_031_sm2) -Education is one of the *_priority_* […] (txt_027_sm2)

Tableau 60 : Quelques exemples d'erreurs de calculs sémantiques

Quant aux aléas de la cognition, on a affaire précisément ici aux aléas accidentels du parcours de la vie. Autrement dit, il s'agit tout simplement d'une erreur non systématique qui relève d'une inattention ponctuelle.

Par ailleurs, pour ce qui est de la cognition, une remarque générale s'impose. Nous reconnaissons avoir observé des chevauchements possibles entre certaines erreurs de calculs sémantiques et certaines erreurs qui sont imputables aux systèmes linguistiques (dits [L(1)], [L(2)] et [L(1.5)]). Il arrive, en effet, que la frontière ne soit pas clairement établie[133]. A ce titre, les erreurs de calculs sémantiques nécessiteront une étude supplémentaire, avec des moyens et une démarche proprement psycholinguistiques, si l'on souhaite (i) corroborer nos résultats et (ii) affiner par la suite nos catégorisations.

En définitive, au vu des possibilités offertes par notre cadre, nous soulignons son caractère principalement explicatif. En effet, Granger & Monfort (1994) soutiennent dans la lignée de Corder (1967, 1973) qu'un projet d'analyse d'erreurs doit comporter au moins deux étapes. La première étant celle de la description purement linguistique et la deuxième celle de l'explication des erreurs. Cependant, comme nous l'avons soutenu au début de la section 8.3.2, malgré les différences observées dans les nombreux schémas d'annotations et les descriptions linguistiques qui y sont faites, beaucoup se rejoignent dans l'étape de l'explication des erreurs. Notre schéma ou grille se veut alors un outil facilitant cette deuxième étape de l'analyse. De plus, un des avantages de ce type de grille est qu'elle est extensible : c'est-à-dire que la hiérarchisation peut être approfondie selon les besoins de l'analyse, de même que les catégories ou niveaux des colonnes se trouvant à la gauche de la grille.

8.4 Limites et perspectives

Cette dernière section de la conclusion aborde les limites et perspectives de notre étude. La section 8.4.1 ne relève pas, à proprement parler, d'observables issus de la présente étude. Elle témoigne plutôt des questions périphériques qui ont accompagné notre réflexion tout au long de ce travail,

[133] Dans l'exemple suivant : *Anyone in *its his right mind would say the Union is doomed […] :* (i) en termes d'accord en nombre « anyone » et le pronom « its » sont parfaitement compatible (ii) en termes d'accord coréférentiel « anyone » en tant que pronom (personnel) indéfini est incompatible avec l'adjectif possessif « its ». Mais dans « *Any one would do, because its use is solely for the purpose of illustration* » 'any one' (dans le sens *n'importe lequel*) qui n'a plus les mêmes propriétés que 'anyone' (dans le sens de *quelqu'un*) se voit correctement lié (tant sémantiquement que formellement) à « its ». Selon donc que l'on se penche sur 'anyone' ou 'its' comme une erreur, la classification ici peut être tant d'ordre sémantique que proprement syntaxique. Bien entendu le choix de « its » peut tout aussi bien être une question d'erreur humaine et donc un aléa de la cognition.

mais qui auraient nécessité plus de temps pour une réflexion et examen critique à part entière. Ainsi, elle aborde la question du défi de la rédaction en langue étrangère, en distinguant clairement entre ce qui relève du linguistique, du culturel et du cognitif. La dernière section, 8.4.2, quant à elle, présente les principales limites que nous avons pu identifier à la fin de notre étude. Ces limites n'invalident aucunement les résultats obtenus, mais nous pensons qu'une meilleure prise en compte de ces éléments – notamment méthodologiques – aurait permis d'éviter un certain nombre des écueils que nous avons rencontrés sans oublier, bien entendu, le fait qu'elle aurait également permis à la fois une meilleure reproductibilité de l'étude et une meilleure généralisation de nos résultats. Enfin, il s'agit tout aussi bien ici d'exposer quelques-unes des questions auxquelles nous n'avons pas pu apporter de réponses définitives et qui demeurent, de ce fait, en suspens. Les limites annoncées et les questions en suspens traduisent, respectivement, à notre sens, des perspectives d'améliorations et des pistes de recherche future.

8.4.1 Rédaction en langue étrangère : un défi multifactoriel

Comme nous l'avons souligné notamment dans le chapitre d'introduction et à travers les différents types d'erreurs relevés dans les chapitres qui ont suivi, une des principales difficultés qu'engendre l'apprentissage d'une langue étrangère est celle de la production effective du discours. C'est-à-dire, la capacité réelle de créer du contenu sémantique répondant à la fois aux règles combinatoires du système linguistique, à la situation d'énonciation et à l'intelligibilité du message.

Dès lors qu'un ou plusieurs éléments ne sont pas suffisamment maîtrisés, l'effort communicatif entier se trouve affecté : on parlerait alors d'erreurs de production écrite ou orale. Ce problème n'est pas anodin et retient l'attention, comme nous l'avons souligné dans la section 1.1, de nombreux spécialistes dans des domaines divers – allant notamment de l'anthropologie culturelle, les sciences de l'éducation, sans oublier, la linguistique bien entendu. Pour tenter donc de saisir la complexité du phénomène, nous le présentons sous trois angles différents : tout d'abord linguistique, et ensuite culturel et cognitif.

Linguistique

L'aspect proprement linguistique a été amplement présenté tout au long de ce travail, notamment dans les chapitres I et V. En effet, une des principales remarques faites lorsque le discours d'un apprenant est erroné est celle de la non-maîtrise linguistique : ce qui signifie de manière générale que les différents moyens mis à disposition de la langue n'ont pas été utilisés ou plus précisément n'ont pas été maîtrisés. Cependant, notons que le terme non-maîtrise linguistique pose un problème

d'interprétation, dans la mesure où ce terme renvoie tant à la grammaire qu'à l'ensemble des moyens qui régissent « une langue » donnée. A ceci s'ajoute l'importante distinction à faire entre connaissances linguistiques et l'utilisation réelle, communément appelées compétence et performance dans la grammaire générative de Chomsky (1965). La première étant la capacité à reconnaitre des énoncés grammaticalement corrects, renvoyant à une connaissance passive des règles de grammaire, tandis que la deuxième renvoie à l'utilisation réelle. Autrement dit, les connaissances grammaticales de l'apprenant ne sont pas toujours synonymes de capacité à produire du discours. De ce fait, les études sur la maturité syntaxique des apprenants constituent une entreprise prégnante et indémodable : tant par le fait que cela occupe une place importante dans l'activité d'enseignement que par le fait qu'elle constitue l'élément clé d'une production intelligible.

Culturel

Hormis l'aspect linguistique qui touche invariablement l'ensemble des locuteurs dits non-natifs sans différenciation de niveau, les problèmes ayant trait à l'aspect culturel sont principalement étudiés chez les apprenants intermédiaires et avancés. Dans l'enseignement supérieur, ceci s'explique par le fait que les productions langagières de ces deux groupes sont envisagées dans une perspective avant tout contextuelle, à la différence des niveaux inférieurs où l'accent est mis sur la correction grammaticale. En effet, dès lors que l'apprenant est capable ou qu'il est demandé à l'apprenant de construire une argumentation développée, quelle que soit la forme retenue, l'accent est mis sur l'acceptabilité de l'ensemble des énoncés comme répondant à une situation de communication donnée.

Cette argumentation, que nous appelons fortuitement ici « discours[134] », n'est pas simplement la somme d'une addition ou l'exploitation de moyens linguistiques d'une langue donnée. Elle puise ces acceptations dans le culturel – qui lui fournit des modalités d'usage en termes de sens et de forme. En ce sens, nous rejoignons Pierre Bourdieu qui voit dans toute utilisation de la langue une dimension singulièrement sociale quand il affirme :

> « Tout acte de parole et, plus généralement, toute action, est une conjoncture, une rencontre de séries causales indépendantes : d'un côté les dispositions, socialement façonnées, de l'habitus linguistique, qui impliquent une certaine propension à parler et à dire des choses déterminées […] et une certaine capacité de parler définie inséparablement comme capacité linguistique d'engendrement infini de discours

[134] Cf. section 2.2 pour une discussion détaillée sur le discours et les pratiques discursives culturelles

> grammaticalement conformes et comme capacité sociale permettant d'utiliser adéquatement cette compétence dans une situation déterminée » (1982 :14)

Cette référence vaut aussi bien pour le locuteur natif que pour le non-natif, vis-à-vis d'une même langue. Ce qui signifie que l'impératif sémantique dans lequel le locuteur natif a socialement baigné fait défaut aux apprenants dès lors qu'ils n'ont pas les mêmes habitus linguistiques. Dans une autre mesure, nous pouvons également relier cette perspective à l'un des postulats majeurs d'Edward Sapir pour qui la langue, son utilisation et la façon de concevoir le réel (comprendre la représentation du monde) sont intimement liées.

> Il est tout à fait illusoire d'imaginer que l'on appréhende la réalité sans passer par l'utilisation de la langue et que la langue n'est qu'un moyen accessoire de résoudre des problèmes spécifiques de communication ou de réflexion. Le fait est que le "monde réel" est dans une large mesure construit inconsciemment sur les habitudes de la langue du groupe. (1929, version réimprimée en 1949 : 69)

Cela signifierait que les différents aspects de l'utilisation de la langue, que ce soit en réception ou en production, sont communément déterminés par la communauté d'appartenance ou d'insertion. Pour toutes ces raisons, il est admis que le discours puise et est considéré comme étant généralement *marqué* par certains aspects fondamentalement ancrés dans le culturel tel un *moule à discours.* C'est ce moule qui permet de produire des formes normalisées ou encore institutionnalisées qui ne sont pas nécessairement existantes ou utilisées de manière identique d'une culture linguistique à une autre. Et c'est bien ce phénomène que nous relevons comme le deuxième problème affectant les productions langagières de nos apprenants de langue étrangère. Toutefois, il conviendrait de signaler ici que des productions d'un niveau intermédiaire ou avancé sont nécessaires pour étudier l'aspect culturel, de manière convenable[135].

Cognitif

Le troisième problème auquel nos apprenants doivent faire face relève de l'ordre cognitif. En effet, nous soutenons que les problèmes cognitifs varient de manière significative entre les individus selon leur niveau de scolarisation et par conséquent leur niveau de littéracie en langue maternelle. Autrement dit, si, par exemple, le principe de concordance des temps n'est pas maîtrisé en langue maternelle, il sera difficile pour l'apprenant de recourir à un usage parfait de celui-ci en classe de langue étrangère. De même, la Linguistic Coding Difference Hypothesis de Ganchow et al. (1998) soutient qu'il existe un lien causal direct entre toute difficulté langagière rencontrée en langue

[135] Nous reviendrons sur ce point dans la section 8.4.2 qui traite des limites de notre étude.

maternelle qui se verrait indubitablement transférée vers la langue étrangère. Nous notons, toutefois, qu'en dépit de l'aspect plausible de cette hypothèse, il manque des contre-tests et des cas d'études pour corroborer cette théorie.

Outre les problèmes de maîtrise de la langue maternelle (avant la maîtrise de la L2), il est judicieux de rappeler que les aptitudes cognitives et la motivation individuelle jouent également un rôle non négligeable dans l'apprentissage d'une langue étrangère. Nous rejoignons, de ce fait, le propos suivant de Hyland (2003).

> No two learners are the same and their different learning backgrounds and personalities will influence how quickly, and how well, they learn to write in a second language. Students obviously bring to the L2 writing class different writing experiences, different aptitudes and levels of motivation; they have varying metacognitive knowledge of their L1 and experience of using it, particularly to write; and they have different characteristics in terms of age, sex, and socioeconomic status. (2003: 32-33)

Mises à part à ces différences cognitives individuelles, nombreux sont ceux qui s'intéressent à l'écriture en langue étrangère comme un objet d'étude à part entière et qui voient dans le procès de la rédaction un défi majeur pour nos apprenants. Et ce, dans la mesure où il est envisagé comme une activité cognitive complexe. Selon Torrance (2006) :

> Text production is therefore necessarily complex, requiring coordination of high-level processes for determining content and structure, and low-level processes associated with producing words on the page. (2006 : 679)

Dans cette optique, les connaissances syntaxiques et lexicales sont considérées comme des processus mineurs, tandis que l'organisation informationnelle créant un ensemble textuel complet serait plus complexe. Ces arguments font écho aux postulats mis en avant par les psycholinguistes Flower et Hayes (1980) et Anderson (1985) qui conçoivent également la rédaction comme une activité hautement complexe. Ces derniers vont jusqu'à détailler les phases obligatoires auxquelles les apprenants-scripteurs doivent se soumettre pour réaliser leur texte[136]. Pour ces derniers, il s'agirait de ce qu'ils appellent le processus de *planning, mise en texte et révision* pour les premiers et *construction, transformation, exécution* pour le deuxième.

Dans une moindre mesure, on pourrait supposer que la conceptualisation d'un fait linguistique qui n'existerait pas en l'état avec les mêmes valeurs en langue maternelle poserait de nombreux problèmes d'utilisation. C'est-à-dire qu'il n'existe pas ou n'a pas un fonctionnement identique

[136] Cf. Fayol (2007) et Piolat & Roussey (1992) sur les processus cognitifs mobilisés pour la rédaction d'un texte.

selon les langues et constitue, de ce fait, un obstacle important pour les apprenants qui ne l'ont pas tel quel dans leurs langues maternelles. On pourrait aussi chercher à faire un parallèle entre le principe de différentiation qui existe à travers les différents temps grammaticaux (*tense*) en anglais ou en français mais qui n'existeraient pas tels quels, par exemple, en mandarin. Nous pouvons supposer que ce type de conceptualisation constituerait un obstacle majeur pour ces apprenants de langue étrangère.

Ces exemples, bien que limités ici, nous permettent de comprendre que la cognition est avant tout liée à la langue maternelle qui régule notre conceptualisation du réel. De ce fait, nous rejoignons le principe de relativisme linguistique qui soutient que la façon d'appréhender le monde – et par extension la façon de l'expliquer – doit être pensée à travers une triade linguistique, culturelle et cognitive. Cela peut expliquer en partie pourquoi certains apprenants ont des difficultés à utiliser une langue étrangère pour exprimer des sentiments, pour effectuer leurs prières voire faire des calculs, puisque ces activités sont avant tout des activités principalement cognitives et intimes – et non arbitrairement réalisables dans tous les systèmes linguistiques à disposition d'un locuteur donné.

En somme, en termes de défi en langue étrangère, il existe de multiples arguments que nous avons choisi de ne pas exposer ici en raison de leurs aspects strictement psychologiques ou par faute de temps. Toutefois, il est à souligner que ces éléments ne sont pas à ignorer mais à prendre en compte – avec leur contexte - selon l'aspect d'acquisition que l'on souhaite étudier. Nous pensons, de ce fait, que les trois aspects mis en avant ci-dessus influent, selon le contexte d'étude, de manière directe sur la façon dont les erreurs sont « interprétées » et ensuite « remédiées ».

8.4.2 Limites et l'étude et pistes pour la suite

Dans cette dernière section, nous passons aux limites de la présente étude en nous intéressant aux questions qui restent en suspens et tout particulièrement à notre réflexion générale sur les limites de quelques-uns de nos choix méthodologiques. Pour les besoins de brièveté, la discussion est divisée en deux parties portant sur (i) le cadre théorique et (ii) le cadre méthodologique. La section se clôt avec des remarques d'ordre général portant sur des perspectives de recherche future.

Le cadre théorique

Pour ce qui est de l'apport de la linguistique systémique fonctionnelle, nos constats sont mitigés. Tout d'abord, le cadre LSF nous a permis d'identifier de façon irréfutable les points et les constructions sémantiques qui font défaut à nos sujets-participants. En effet, bien que le syntagme

nominal revête de manière générale un certain nombre d'écueils pour nos apprenants (cf. section 5.1.2), nous savons désormais que les syntagmes en position d'objet ou de complément (c'est-à-dire, dans le rôle de participant 2, cf. section 5.1.4) sont deux fois plus susceptibles de conduire à des erreurs grammaticales que ceux en position de sujet. Le cadre LSF a également permis de mettre en avant le fait que les circonstances erronées – tout particulièrement celles dites de « manière » et de « location » présentent la même fréquence d'occurrences erronées que les syntagmes nominaux en position d'objet. En effet, parmi les nombreux types de circonstances, les deux types cités ci-dessus sont plus susceptibles de conduire à des agrammaticalités tandis que celles dites « d'angle » conduisent plus à des questions d'acceptabilité textuelle (cf. section 6.2.2).

Mises à part ces considérations proprement grammaticales, nous avons également pu identifier un autre trait qui semble constituer un véritable problème pour nos apprenants : à savoir le thème interpersonnel dit « commentaire ». Cette valeur proprement textuelle assure une certaine position argumentative dans le discours. En effet, cette valeur permet d'ajouter un positionnement personnel (cf. section 3.2.5 et 6.2.4), en adjoignant des précisions principalement évaluatives sur le contenu discursif. Mais comme nous l'avons vu aussi bien dans le chapitre V que dans le chapitre VI, cette valeur est une des plus problématiques qui conduit irrémédiablement tant à des problèmes d'agrammaticalité qu'à des problèmes d'acceptabilité textuelle. Nous soulignons donc ici que le cadre LSF a facilité, entre autres, ces observations qui n'auraient été possibles ni avec un schéma d'annotation d'erreur traditionnel ni avec un schéma issu de la grammaire traditionnelle.

Hormis les précisions signalées ci-dessus, nous soulignons deux de ses principales limites dans notre étude : la première limite est d'ordre pratique et la deuxième limite est d'ordre conceptuel. En effet, lorsque nous avons procédé à des comparaisons de nos résultats avec ceux des études antérieures, nous avons mis en avant le fait qu'un flou terminologique existait entre les projets et les schémas d'annotations différents – par voie de conséquence ce flou terminologique obstruait en quelque sorte le processus de comparaison. Nous nous sommes alors rendu compte que la terminologie proprement LSF risquait d'ajouter, à son tour, au flou terminologique que nous avons préalablement dénoncé. Nous reconnaissons de ce fait que de la même manière que le terme « erreur lexicale » renvoie à des acceptations différentes dans la tradition générale de l'analyse des erreurs, le terme « erreur de procès », par exemple, avec ses six sous-types différents risque de présenter un certain nombre d'obstacles pour celui qui n'est pas proprement initié à la terminologie LSF et qui voudrait procéder à des comparaisons ou à la reproduction de notre étude.

La deuxième limite du cadre de la linguistique systémique fonctionnelle réside dans le fait que cette dernière n'a pas de dimension proprement cognitive ou psycholinguistique : ce qui fait que le cadre lui-même offre des moyens considérables d'analyses concrètes mais ne fournit pas pour autant l'ensemble des moyens nécessaires pour la phase de l'interprétation. Et comme nous l'avons soutenu dans les sections 8.3.2 et 8.4.1, si l'on souhaite obtenir une vue d'ensemble sur le processus de l'acquisition d'une langue étrangère, la dimension cognitive doit être considérée comme un aspect à part entière dans toute analyse des productions en langue étrangère – sans oublier, bien entendu, les facteurs proprement linguistiques et culturels.

Le cadre méthodologique

La plupart des limites que nous avons pu identifier dans la présente étude concerne nos choix méthodologiques. Cela étant, nous les passons succinctement en revue. Tout d'abord, nous nous sommes demandé si les résultats obtenus principalement dans les chapitres V et VI auraient été différents ou plus significatifs si nous avions procédé différemment. Nous présentons ci-après quelques-unes des questions que nous nous sommes posé, une fois que l'analyse a été achevée.

- Est-ce qu'une approche axée davantage sur la dimension qualitative (c'est-à-dire, avec moins de sujets-participants mais plus de productions par participant individuel) aurait permis de mieux rendre compte et d'étudier les différents types d'erreurs que nous avons observés ? Aurait-elle permis une meilleure précision sur le processus acquisitionnel des différents éléments linguistiques à l'étude ?
- Est-ce que les trois profils linguistiques (à savoir, E+CLR, F-CLR et F+CLR)[137] auraient pu être mieux établis ? Est-ce que des tests standardisés, du type « Oxford placement test », « First Certificate in English (FCE) », « TOEIC[138] » et ainsi de suite, auraient permis de le faire ? De plus, est-ce que ces profils auraient dû être sous-divisés pour mieux repartir des participants en sous-groupes homogènes ? Et est-ce qu'une meilleure répartition aurait conduit à des résultats différents ?
- Enfin, est-ce que l'on aurait gagné à affiner les profils F-CLR et F+CLR, en demandant aux sujets-participants des précisions non seulement sur la durée de leur séjour dans des pays anglophones, mais la date ou leur âge au moment où le séjour a été effectué – de manière à établir toute corrélation possible entre un séjour effectué, par exemple, à 4 ans et un autre effectué à 17 ans ?

[137] Cf. Annexe (A2)
[138] Test of English for International Communication *(*TOEIC*)*

Ces questions sur le profil ou le « profilage » des participants méritent réflexion. Et si nous devions refaire notre étude, ces questions seront intégrées à la phase de conception méthodologique. En effet, bien que ces questions n'invalident pas notre étude, la précision apportée par ces facteurs aurait permis d'avoir des « profils historiques langagiers » beaucoup plus précis. Ces précisions auraient également permis, à notre avis, de mieux cibler et d'étudier l'évolution des erreurs à l'intérieur des différents sous-groupes et cela aurait (peut-être) facilité l'émergence de davantage de traits distinctifs que l'on aurait pu exploiter directement en cours de langue.

Le dernier point sur le choix de participants concerne la variable temporelle. Nous aurions aimé pouvoir étudier des productions écrites des apprenants en anglais langue étrangère sur une période plus longue : à savoir pendant 36 mois (de la rentrée des participants en première année à l'université jusqu'à la fin de leur troisième année, de façon à suivre l'évolution non seulement des erreurs mais de leur compétence linguistique générale). Malheureusement, nous n'avons pas été autorisé à les étudier que pendant une période restreinte : c'est-à-dire, pendant deux semestres[139]. Cette piste sera donc à poursuivre dans un projet ultérieur.

Mises à part les réflexions sur les sujets-participants de notre étude, nous passons maintenant à la dernière limite méthodologique que nous avons identifiée dans notre étude. Un élément qui constitue un point central à notre travail – tout en recouvrant une certaine limite – est le schéma d'annotation d'erreur intégré au logiciel UAM CorpusTool. En effet, il nous parait judicieux de souligner ici que le logiciel et son schéma d'annotation ont été modifiés par son concepteur au cours de notre travail de doctorat. Cela signifie donc que le schéma dans notre travail n'est pas « à jour ». Mais bien que cela ne soit qu'un aspect mineur, nous tenons à expliquer quelques-unes des modifications qui ont été apportées par le développeur et notre position par rapport à ces multiples mises à jour.

- Tout d'abord, dans la nouvelle version du logiciel le développeur propose désormais deux schémas d'annotation : une version simplifiée et une version longue. La version simplifiée comporte une cinquantaine d'étiquetages par rapport au schéma de l'ancienne version (version 2.8) qui proposait trois fois plus d'étiquetages. Il y a toujours les six principales catégories : à savoir lexicales, grammaticales, pragmatiques, de mise en phrase (phrasing, en anglais) et incodables. Ce qui change est que beaucoup de catégories n'ont plus que deux niveaux de profondeur par rapport aux quatre niveaux de profondeur initialement proposés.

[139] Le refus venait de la part, non pas de l'administration universitaire, mais des enseignants.

- Mais sans nous attarder davantage sur le schéma simplifié, intéressons-nous au niveau schéma en version longue (version 3.1). La première chose que l'on remarque est que les erreurs lexicales sont divisées en deux sous-catégories : la première sous-catégorie renvoie plus au moins à celle que nous avons détaillée dans les chapitres IV et V ; et la deuxième sous-catégorie regroupe les erreurs de choix lexicaux qui étaient reliés à la nature ou la classe grammaticale dans la version 2.8. Ce premier changement nous « conforte » en quelque sorte. Dans la section 8.3.1.2 nous avons mis en avant le fait que les erreurs lexicales issues du schéma UAM n'étaient pas « signalées » de la même manière que les erreurs lexicales des quatre autres études avec lesquelles nous avons procédé aux comparaisons. Cela nous avait plus au moins contrait à (i) adopter, « en cours de route », la définition de Granger et Monfort (1994) portant sur les erreurs lexicales et (ii) de réajuster, par voie de conséquence, l'ensemble des différentes classifications et leurs calculs statistiques. Le premier changement du nouveau schéma d'UAM va donc dans le sens de l'harmonisation des schémas d'annotation que nous avons affirmé vouloir voir se matérialiser dans la section 8.1.
- Un autre point qui mérite d'être signalé est celui de l'étiquetage « cohesion error » de la version 2.8. En effet, dès que nous avons commencé le travail d'annotation, nous avons ajouté deux sous-catégories à cet étiquetage qui n'en comportait aucun : les deux ajouts concernaient (i) un étiquetage pour des problèmes de saillance entre référents et (ii) un autre pour des problèmes d'incompatibilité entre référents. Nous remarquons que le nouveau schéma d'UAM comporte désormais trois sous-catégories étiquetées (i) « reference error » ; (ii) « ellipse error » ; et (iii) « substitution error ». A ce constat s'ajoute également le fait que nous sommes arrivés à développer deux catégories identiques, en même temps et d'une manière similaire que le développeur lui-même. En effet, nous avions ajouté deux sous-catégories aux erreurs dites de cohérence vis-à-vis du contexte textuel : l'une portant sur l'incohérence du modal auxiliaire et l'autre portant sur l'incohérence du connecteur (et le développeur a donc fait de même avec plus au moins les mêmes noms).
- Toutefois les quelques ajouts supplémentaires ou les changements d'étiquetage d'appartenance (ou de dépendance) – notamment par rapport aux anciens et nouveaux étiquetages et les nœuds ou les catégories auxquelles ils appartiennent – font que le schéma que nous avons employé dans la présente étude n'est ni complètement identique à la version du développeur en 2012 (version 2.8) ni celle d'aujourd'hui (version 3.1). Ces différences peuvent en principe constituer un problème pour la comparabilité et notamment pour la reproductibilité de l'étude. Mais comme nous l'avons vu dans la section 8.3.2, le fait que

des étiquetages de deux schémas d'annotations différents ont des noms ou des dépendances différentes ne veut pas dire que l'on ne peut pas procéder à une comparaison objective des deux études. Dans ces cas précis, il est tout simplement question de procéder à une analyse minutieuse de l'acceptation précise d'un terme ou d'un étiquetage et de trouver son correspondant dans le schéma que l'on souhaite comparer.

En guise de conclusion, cette étude a permis d'apporter un certain nombre d'éclaircissements aux questions que nous nous sommes posés en début de ce travail de doctorat (cf. section 0.1). Elle a, par exemple, confirmé des hypothèses émises sur (i) l'existence d'erreurs de type proprement textuel (bien que nous reconnaissions que la frontière permettant de distinguer entre erreurs d'acceptabilité textuelle n'est pas toujours fixe); (ii) et par voie de conséquence, le fait que toutes les erreurs ne peuvent s'expliquer en termes de connaissances grammaticales ou lexico-grammaticales acquises par un apprenant ; (iii) l'existence d'une frontière distincte entre les erreurs lexicales individuelles et les erreurs lexicales multi-mots (dites aussi, des erreurs de phraséologie lexicale) – sans oublier, bien entendu, que nous avons apporté un schéma facilitant la première étape de distinction et d'explication pour l'ensemble des erreurs que nous avons examiné tout au long de ce travail ; et enfin (iv) le fait que l'ajout d'un cadre linguistique précis (comme nous l'avons fait avec la LSF) pourrait apporter son « lot de contribution » à l'avancement du cadre d'analyse d'erreurs.

Mais cette étude a également ouvert la voie à plusieurs questions que nous ne nous étions pas posé en début de ce travail. Ces questions nécessiteront, pour la plupart, une étude supplémentaire, ne serait-ce que pour infirmer ou confirmer les premières hypothèses émises et les premiers résultats obtenus dans le présent travail. A ce titre, notre étude constitue à nos yeux le premier tremplin d'une entreprise qui s'annonce passionnante.

Bibliographie

Acevedo, C. & Rose, D. (2007) Reading (and writing) to learn in the middle years of schooling, Pen 157, Sydney : Primary English Teaching Association, pp. 1-8

Adam, J-M. (1990) Eléments de linguistique textuelle : théorie et pratique de l'analyse textuelle. Paris : Mardaga

Adam, J-M. (1991) Langue et littérature : Analyses pragmatiques et textuelles. Paris : Hachette.

Adam, J-M. (1993) Le texte et ses composantes. In Semen, Revue de sémio-linguistique des textes et discours

Adam, J-M. (2003) Entre la phrase et le texte : la période et la séquence comme niveaux intermédiaire de cohésion. In Québec français, n° 128, pp. 51-54. URL : http://id.erudit.org/iderudit/55780ac

Adam, J-M. (2004) Linguistique textuelle. Des genres de discours aux textes. Une introduction méthodique à l'analyse textuelle des discours. Paris : Nathan

Adam, J-M. (2005) La linguistique textuelle. Introduction à l'analyse textuelle des discours. Paris : Armand Colin

Adolphs, S & Lin, P.M.S. (2011) Corpus Linguistics, In Simpson, J. (éds). The Routledge Handbook of Applied Linguistics. Abingdon : Routledge

Albert, C., Garnier, M., Rykner, A., & Saint-Dizier, P. (2009) Eléments de stratégie de correction automatique de textes : le cas des francophones s'exprimant en anglais. In Congrès de l'Acfas, Ottawa, Presses de Université du Québec, pp. 34-42

Alderson, J., Clapham, C., & Steel, D. (1997) Metalinguistic knowledge, language aptitude and language proficiency. Language Teaching Research 1(2), pp. 93-121

Anctil, D. (2010) L'erreur lexicale au secondaire : analyse d'erreurs lexicales d'élèves de 3e secondaire et description du rapport à l'erreur lexicale d'enseignants de français. Thèse de doctorat, Université de Montréal

Andersen, Ø. (2011) Semi-automatic ESOL error annotation. English Profile Journal, 2. Cambridge University Press

Anthony, L. (2013a). A critical look at software tools in corpus linguistics. Linguistics Research 30 (2), pp. 141-161

Anthony, L. (2013b) Developing AntConc for a new generation of corpus linguists, Proceedings of the Corpus Linguistics Conference (CL 2013), Lancaster University, pp. 14-16

Arbach, N. & Ali, S. (2013) Aspects théoriques et méthodologiques de la représentativité des

corpus, Corela, HS-13 | consulté le 21 juillet 2015. URL : http://corela.revues.org/3029

Arnaud, P. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In Culhane, T., Klein Bradley, C., & Stevenson, D. (éds). Practice and problems in language testing: papers from the International Symposium on Language Testing. University of Essex, pp. 14-28

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), pp. 555–596

Authier, J. & Meunier, A. (1972) Norme, grammaticalité et niveaux de langue In Langue française, 16, pp. 49-62.

Bache, C. (2010) Hjelmslev's Glossematics: A source of inspiration to Systemic Functional Linguistics? Journal of Pragmatics 42, pp. 2562–2578

Bachman, L. F. (1990) Fundamental Considerations in Language Testing. Oxford University Press

Banks, D. (2005) Introduction à la linguistique systémique de l'anglais, Paris : L'Harmattan.

Banks, D. (2010) The interpersonal metafunction in French from a Systemic Functional Perspective. Languages Sciences, vol 32, 3, pp. 395-409

Barbier, M-L. (2003) Écrire en L2 : bilan et perspectives des recherches. Arobase, www.arobase.to, volume 1-2, pp. 6-21

Barrett, N.E & Chen, L. (2011) English Article Errors in Taiwanese College Student's EFL Writing. Computational Linguistics and Chinese Language Processing, 16, 3-4, pp. 1-20

Bateman, J. A. (1990) From Systemic-Functional Grammar to Systemic-Functional Text Generation: Escalating the Exchange, ISI/RR-89-220. University of Southern California

Bateman, J. A. (1995) KPML: The KOMET-Penman multilingual resource development environment. In Proceedings of the Fifth European Workshop on Natural Language Generation, pp. 219-222

Bateman, J. A. (1997) Enabling technology for multilingual natural language generation: the KPML development environment Natural Language Engineering, 3(1), pp. 15-55.

Bateman, J. A., Matthiessen, C.M.I.M., & Zeng, L. (1999) Multilingual natural language generation for multilingual software: a functional linguistic approach. Applied Artificial Intelligence, 13 (6), pp. 607-639

Beaugrande, R. de (1998) On 'usefulness' and 'validity' in the theory and practice of linguistics: A riposte to H.G. Widdowson. Functions of Language 5 (1), pp. 87-98.

Birner, B. & Ward, G. (2004) Information Structure and Non-canonical Syntax. In Horn, L. R. & Ward, G. (éds) The Handbook of Pragmatics, Oxford: Blackwell, pp. 153-174,

Biber, D. (1993) Representativeness in corpus design, Literary and linguistic computing, 8 (4), pp. 243–257

Biber, D. (2006) University language: A corpus-based study of spoken and written registers. Amsterdam : John Benjamins

Biber, D & Susan, C. (1999) "Lexical Bundles in Conversation and Academic Prose." *Out of Corpora: Studies in Honor of Stig Johansson.* Eds. Hilde Hasselgard and Signe Oksfjell. Amsterdam: Rodopi. pp.181-9.

Biber, D., Connor, U., & Upton, T. (2007) Discourse on the move: Using corpus analysis to describe discourse structure. Amsterdam : John Benjamins.

Bitchener, J., Young, S & Cameron, D (2005) The Effect of Different Types of Corrective Feedback on ESL Student Writing. Journal of Second Language Writing

Bloor, T. & Bloor, M. (2004) The Functional Analysis of English : A Hallidayan Approach. London : Edward Arnold

Boulton, A. 1998. L'acquisition du lexique en langue étrangère. In: Actes du 26ème Congrès de l'UPLEGESS, pp. 77-87

Bourdieu, P. (1982) Ce que parler veut dire : l'économie des échanges linguistiques. Paris : Fayard

Bruce, R. & Wiebe, J. (1999) Recognizing subjectivity: A case study in manual tagging. Natural Language Engineering, 5(2), pp. 187–205

Busch-Lauer, I. (2002) Technical vs. Academic Writing in English – Any Difference for Non-native Writers? ASp, Rédactologie, pp. 37-38.

Bussmann, H. (1996) Routledge Dictionary of Language and Linguistics. Londres : Taylor & Francis.

Butler, C. (2003) Structure and Function: From clause to discourse and beyond. John Benjamins Publishing

Cabassut, E. (2003). De la linguistique des fautes à une didactique multilingue. Thèse de doctorat, Université Marc Bloch - Strasbourg II.

Caffarel-Cayron, A. (2009). La répresentation grammaticale de l'expérience: transitivité et agence. In Eason, S., Ormrod, J. & Banks, D (éds). La linguistique systémique fonctionnelle et la langue française, Paris : L'Harmattan, pp. 67-88

Carter-Thomas, S. (1999a) Thematic networks and text types. In ASp, la revue de Geras, 23/26, pp. 139-148

Carter-Thomas, S. (1999b) L'organisation thématique et ses consequences sur la clarté d'un texte. In Le corps et le Langage, l'Harmattan, pp. 121-137

Carter-Thomas, S. (1999c) La stratégie thématique : son importance dans l'analyse textuelle. In 5ème journée de la Formation Doctorale de Linguistique Générale et Appliquée, Université Paris Descartes

Carter-Thomas, S. (1999d) Erreurs locales et erreurs globales : une contribution à l'analyse textuelle de l'anglais scientifique. In Actes de la 7ème journée ERLA/GLAT, pp. 267-280

Carter-Thomas, S. (2000) La cohérence textuelle : pour une nouvelle pédagogie de l'écrit. Coll. Langue et parole. Paris : L'Harmattan.

Carter-Thomas, S. (2009a) Teaching coherence through genre. In Alamargot, D., Bouchand, J., Lambert, E., Millogo, V., & Beaudet, C. (éds)., Actes du colloque « De la France au Québec. L'écriture dans tous ces états » Poitiers. [http://www.ecritfrancequebec2008.org/]

Carter-Thomas, S. (2009b) Texte et contexte : pour une approche fonctionnelle et empirique. Mémoire de synthèse d'HDR, Université Sorbonne Nouvelle – Paris 3

Carter-Thomas, S & Rowley-Jolivet E. (2001) Syntactic differences in oral and written scientific discourse: the role of information structure, *ASp*, Vol. 31, pp. 19-37.

Charles, M., Percorari, D. & Hunston, S. (2009) Academic Writing: At The Interface Of Corpus and Discourse. Londres : Continuum International Publishing Group

Charolles, M. (1978) Introduction aux problèmes de la cohérence des textes. In. Langue française, n°38, pp. 7-41

Charolles, M. (1989) Problèmes de la cohérence textuelle. In Charolles, M., Halté, J.F., Masseron, C. & Petitjean, A (éds). Pour une didactique de l'écriture, collection "Didactique des Textes", Université de Metz, pp. 9-49

Charolles, M (2009). Les cadres de discours et leurs frontières. In Delomier, D. & Morel, M-A. Frontières : du linguistique au sémiotique, Lambert-Lucas, pp. 143-162

Charolles, M., Petöfi, J. & Sözer, E. (1986) Research in text connexity and text coherence, Papiere zur textlinguistik. Band 53,1, Hamburg: Buske

Christie, F. (1991) Literacy in Australia, Annual Review of Applied Linguistics 12, pp. 142-155

Christie, F. (2004) Systemic functional linguistics and a theory of language in education. Systemic Functional Linguistics in Action, A Special Edition of the Journal of English Language, Literatures in English and Cultural Studies, pp. 13–40

Chomsky, N. (1965) Aspects of the Theory of Syntax. Cambridge, MA : MIT Press.

Chuang, F-Y. & Nesi, H. (2006) An analysis of formal errors in a corpus of l2 English produced by Chinese students. Corpora, 1 (2), pp. 251-271

Clyne, M. (1981) Culture and Discourse Structure. Journal of Pragmatics, 5(1), pp. 61–66.

Clyne, M. (1987) Cultural differences in the organisation of academic texts: English and German. Journal of Pragmatics, 11, pp. 211-247

Clyne, M. (2002) Contrastive discourse studies. Cahiers de Praxématique, 38, pp. 59-84.

Clyne, M. (2003) Dynamics of language contact: English and immigrant languages. Cambridge: Cambridge University Press.

Coffin, C. & Donohue, J. P. (2012) Academic Literacies and systemic functional linguistics: How do they relate? Journal of English for Academic Purposes 11, pp. 64–75

Combettes, B. (1983) Pour une grammaire textuelle : La progression thématique. De Boeck-Duculot

Combettes, B. & Tomassone, R. (1988) Le texte Informatif : Aspects linguistiques. De Boeck-Université, Coll. Prisme, Série Problématiques

Connelly, M. E. (135) A remedial drill for Correcting the language errors of children. Mémoire de master, Université de Boston

Connor, U. (2002) New Directions in Contrastive Rhetoric, TESOL Quarterly, 36, (4), pp. 493-510

Connor, U. (2004) Intercultural rhetoric research: Beyond texts. Journal of English for Academic Purposes 3, pp. 291–304

Connor, U. & Kaplan, R. (1987) Writing across Languages: Analysis of L2 Text. Boston, MA : Addison Wesley

Corder, S.P. (1967) The Significance of Learners` Errors. International Review of Applied Linguistics, 5, pp. 161-169. Traduit et publié en français en (1980a) Que signifient les erreurs des apprenants ? In Langages, 14e année, n°57, pp. 9-15

Corder, S. P. (1971a) Le rôle de l'analyse systématique des erreurs en linguistique appliquée. In Bulletin CILA (Commission interuniversitaire suisse de linguistique appliquée) Bulletin VALS-ASLA, vol. 14, pp. 6-15

Corder, S. P. (1971b). Idiosyncratic dialects and error analysis. IRAL, 9, (2), pp. 147-160. Traduit et publié en français (1980b) Dialectes idiosyncrasiques et analyse d'erreurs. In Langages, 14e année, n°57, pp. 17-28.

Corder, S.P. (1973) Introducing Applied Linguistics. Royaume Uni : Penguin Education

Corder (1980a) Que signifient les erreurs des apprenants ? In Langages, 14e année, n°57, pp. 9-15

Corder (1980b) Dialectes idiosyncrasiques et analyse d'erreurs. In Langages, 14^{e} année, n°57, pp. 17-28

Cowie A.P. (1998) (éds). Phraseology: Theory, Analysis and Applications. Oxford: Oxford University Press.

Dabène, L. (1992) Le développement de la conscience métalinguistique: un objectif commun pour l'enseignement de la langue maternelle et des langues étrangères, dans Repères, n°6, pp. 13-23

Dagneaux, E., Denness, S. & Granger, S. (1998) Computer-aided error analysis. System 26, pp. 163-174

De Felice, R. (2008). Automatic error detection in non-native English. Thèse de doctorat, University d'Oxford

Degand, L., & Hadermann, P. (2009) Structure narrative et connecteurs temporels en français langue seconde. Représentations du sens linguistique IV (Helsinki, du 28 au 30/05/2008). In Havu, E., Harma, J., Helkkula, M., Larjavaara, M. & Tuomarla, U (éds) La langue en contexte. Actes du colloque « Représentations du sens linguistique IV » Société néophilologique : Helsinki, pp. 19-34

Diaz-Negrillo, A. & Fernandez-Dominguez, J. (2006) Error tagging systems for learner corpora, RESLA 19, pp. 83-102

Dörnyei, Z. (1998). Motivation in second and foreign language learning. Language Teaching, 31, pp. 117-135

Drury, H. (1991) The use of systemic linguistics to describe student summaries at university level. In Ventola, E. (éds). Functional and Systemic Linguistics: Approaches and Uses. Berlin, New York: Mouton de Gruyter, pp. 431-456

Dumont, B. (1989) Questionnements et interprétation des erreurs en mathématiques : élaboration de modèles pour la compréhension des comportements de réponse et la construction d'outils pédagogiques à support technologiques. Thèse de doctorat d'état, Université Paris 7.

Eggins, S. (2004) An Introduction to Systemic Functional Linguistics. Londres & New York : Continuum

Ellis, R & Barkhuizen, G (2005) Analysing learner language. Oxford : Oxford University Press.

Ellis, R. (2006) Current Issues in the teaching of Grammar: An SLA Perspective. TESOL QUARTERLY, vol. 40, (1)

Fawcett, R. (2000) A Theory of Syntax for Systemic Functional Linguistics. John Benjamins

Fawcett, R. (2005) Invitation to Systemic Functional Linguistics. The Cardiff Grammar as an extension and simplification of Halliday's Systemic Functional Grammar (second edition). Equinox Publishing

Fawcett, R. (2004) Systemic Functional Grammar as a formal model of language: a micro-grammar for some central elements of the English clause. Article non publié

Fawcett, R. & Tucker, G. (1990) Demonstration of GENESYS: a very large semantically based Systemic Functional Grammar. In Proceedings of the 13th Int. Conf. on Computational Linguistics (COLING '90)

Fayol, M. (2007) La production des textes et son apprentissage. In Actes de colloques « Les journées de l'Observatoire National de la Lecture (ONL) Écrire des textes, l'apprentissage et le plaisir »

Fillmore, Ch. J. (1992) “Corpus linguistics” vs “Computer-aided armchair linguistics”, Directions in Corpus Linguistics, pp. 35-60. La Hague : Mouton de Gruyter

Fitikides, T. J. (2003) Common Mistakes in English with Exercises. Harlow : Longman

Fontaine, L. (2012). Analysing English Grammar: A Systemic Functional Introduction. Cambridge : Cambridge University Press

Fontaine, L. (2013) Choice in contemporary systemic functional theory. In Fontaine, L., Bartlett, T. & O’Grady, G. (éds). Systemic Functional Linguistics: Exploring Choice. Cambridge : Cambridge University Press.

Fontaine, L. & Kodratoff, Y. (2002) La notion de « concept » dans les textes spécialisés : une étude comparative entre la progression thématique et la texture des concepts. In ASp, Rédactologie-Situations d’apprentissage, n° 37-38

Forsyth, H. (2014). The Influence of L2 Transfer on L3 English Written Production in a Bilingual German/Italian Population: A Study of Syntactic Errors. Open Journal of Modern Linguistics, 4, pp. 429-456.

Fort, K. (2012) Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. Thèse de doctorat, Université Paris 13.

Frei, H. (1929) La grammaire des fautes, ré-édition Ennoïa, 2004

Frunza, O. Et Inkpen, D. (2007) A tool for detecting French-English cognates and false friends, TALN 2007

Ganschow, L., Sparks, R. L. & Javorsky, J (1998) Foreign Language Learning Difficulties: An Historical Perspective. Journal of Learning Disabilities, Vol. 31 (3), pp. 248-258

Galtung, J. (1981) Structure, Culture and Intellectual Style. Social Science Formation 20, pp. 817-856

Gledhill, C. (2011). La grammaire générative : une introduction critique et une confrontation avec le modèle systémique fonctionnel. Amadis 9, pp. 341-374

Gombert, J-É., (1996) Activités métalinguistiques et acquisition d'une langue, Acquisition et interaction en langue étrangère [En ligne], 8 | 1, mis en ligne le 05 décembre 2011, consulté le 23 juillet 2015. URL : http://aile.revues.org/1224

Gonzaga, J. J. (2011). Intricate Cases in Clauses In SFG Concerning The Grammar Of Brazilian Portuguese. Thèse de doctorat, Université Fédérale de Santa Catarina

Granger S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer K., Altenberg B. and Johansson M. (éds). Languages in Contrast. Text-based cross-linguistic studies. Lund Studies in English 88, Lund: Lund University Press pp. 37-51

Granger, S. (2002) A Bird's Eye View Of Learner Corpus Research. In Granger, S., Hung, J. & Petch-Tyson, S. (éds). Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, Amsterdam & Philadelphia : Benjamins, pp. 3-33

Granger, S. (2003) Error-tagged learner corpora and CALL: A promising synergy, CALICO Journal, 20(3) pp. 465–80

Granger, S. (2004) Computer learner corpus research: current status and future prospects. In Connor, U., & Upton, T. A. (éds). Applied corpus linguistics: A multidimensional perspective. Amsterdam : Rodopi Publishers, pp. 123-145

Granger S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) Corpus Linguistics. An International Handbook, 1, Berlin & New York : Walter de Gruyter, pp. 259-275

Granger, S. & Monfort, G. (1994) La description de la compétence lexicale en langue étrangère : perspectives méthodologiques, Acquisition et interaction en langue étrangère, 3, pp. 55-75

Granger, S., Vandeventer A. & Hamel M-J. (2001). Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL. Linguistique de Corpus. Special issue of Traitement automatique des langues 42 (2), pp. 609-621

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In Granger, S. & Meunier, F. (éds). Phraseology: An Interdisciplinary Perspectives. Amsterdam & Philadelphia : Benjamins, pp. 27-49

Gries, S. Th. (2009) What is Corpus linguistics? Language and Linguistics Compass, 3, Blackwell Publishing, pp. 1-17

Gries, S. Th. (2010) Corpus linguistics and theoretical linguistics, A love–hate relationship? Not necessarily… In International Journal of Corpus Linguistics 15 (3), pp. 327–343

Grossmann, F., Paveau M-A & Gérand, P. (2005) Didactique du lexique: langue, cognition, discours. Grenoble : ELLUG.

Gut, U. & Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In Proceedings of the Speech Prosody, Nara, Japon, pp. 565–568

Habert, B. (2004) Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs, Revue française de linguistique appliquée (Vol. IX) (1), pp. 5-24 URL : www.cairn.info/revue-francaise-de-linguistique-appliquee-2004-1-page-5.htm.

Habert, B., Nazarenko, A., & Salem, A. (1997) Les linguistiques de corpus. Paris : Armand Colin

Halliday, M.A.K. (1994) Introduction to Functional Grammar. Londres & New York : Arnold.

Halliday, M.A.K. (2002) Computing Meanings: Some Reflections on Past Experiences and Present Prospects. In Huang, G & Wang, Z (éds). Discourse and Language Functions. Shangai : Foreign Language Teaching Press, pp. 3-25

Halliday, M.A.K. (2003) On The "Architecture" of Human Language. In On Language and Linguistics, “Volume 3 in the Collected Works of M.A.K. Halliday”. Londres & New York: Continuum. pp. 1-29

Halliday, M.A.K. & Hasan, R. (1976) Cohesion in English, Longman.

Halliday, M.A.K. & Hasan, R. (1985) Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective. Oxford : Oxford University Press.

Halliday, M.A.K. & Matthiessen, C.M.I.M. (1999) Construing Experience through Meaning: A Language Based Approach to Cognition. Londres & New York : Cassel

Halliday, M.A.K. & Matthiessen, C.M.I.M. (2004) Introduction to Functional Grammar. 3ème édition, Londres : Edward Arnold.

Hamilton, C. (2011) L’apport de l’analyse thématique et informationnelle dans l’étude des problèmes de textes écrits en langue étrangère : cas d’étudiants d’origine chinoise inscrits à Université Paris Descartes. Mémoire de master II, Université Paris Descartes

Hasan, R. 2009. The Place of Context in a Systemic Functional Model. In M.A.K. Halliday & J.J. Webster (éds). Continuum Companion to Systemic Functional Linguistics. Londres & New York : Continuum

Hasselgard, H. & Johansson, S. (2011) Learner corpora and contrastive interlanguage analysis. In Meunier F., De Cock S., Gilquin G. & Paquot M. (éds). A Taste for Corpora. In honour of Sylviane Granger. Amsterdam & Philadelphia: Benjamins, pp. 33-62.

Hemchua, S. & Schmitt, N. (2006) An analysis of lexical errors in the English compositions of Thai learners, Prospect vol. 21, (3) pp. 3-25

Hernandez, G., Duarte, Gloria., & Espejo, M. (2013) La maturité syntaxique en espagnol des étudiants du premier et dernier semestre, de la licence en espagnol, anglais et français de Université de La Salle. Lenguaje, vol.41, (1), pp. 17-34

Hidden, M-O. (2008) Variabilité culturelle des genres et didactique de la production écrite - Analyse longitudinale de textes narratifs et argumentatifs rédigés par des apprenants de français langue étrangère. Thèse de doctorat, Université Sorbonne Nouvelle

Hudson, R. (2001). Grammar Teaching and Writing Skills: the Research Evidence. Syntax in the Schools, 17, pp. 1-6

Hudson, R. (2009). Measuring maturity. In Beard, R., Myhill, D., Nystrand, M., & Riley, J. (éds). The SAGE handbook of writing development. Londres : SAGE Publications, pp. 349-363

Hunston, S. (2006) Corpus Linguistics. In Brown, K. (éds). The Encyclopedia of Language and Linguistics 2ème édition. Oxford : Elsevier, pp. 234-248

Hunt, K. W. (1965) A synopsis of clause-to-sentence length factors. The English Journal. Vol 54, 4, pp. 300–309

Huot, D. & Schmidt, R. (1996) Conscience et activité métalinguistique. Quelques points de rencontre, Acquisition et interaction en langue étrangère [En ligne], 8 | mis en ligne le 05 décembre 2011, consulté le 23 juillet 2015. URL : http://aile.revues.org/123

Hyland, K. (2003) Second Language Writing. Cambridge : Cambridge University Press

Hyland, K. (2009) Academic Discourse. Londres : Continuum

James, C. (1998) Errors in language learning and use: Exploring Error Analysis. Londres : Longman

Jin, S. (2000) Mother tongue reliance and avoidance strategies in second language learning: a study of English majors at four tertiary institutions in P.R. China. Thèse de doctorat, Université de Hong Kong

Johnson, K. & Johnson, H. (1999) (éds). Encyclopedic Dictionary of Applied Linguistics. Blackwell Publishing, Blackwell Reference Online.

Judet de la Combe, P. & Wismann, H. (2004) L'avenir des langues : Repenser les humanités. Paris : Editions du Cerf

Kaplan, R. (1966) Cultural Thought Patterns in Inter-cultural Education. In Language Learning, 16, pp. 1-20.

Kaplan, R. B. (1971) Composition at the Advanced ESL Level: A Teacher's Guide to Connected Paragraph Construction for Advanced-Level Foreign Students. The English Record, vol. 21 (4) pp. 53-64

Kawecki, R (2009) Un corpus antillais d'apprenants de français. In Williams, G (dir.) Texte et Corpus, n°4, Actes des 6ème Journées de la linguistique de Corpus, pp. 123-134

Kramsch, C. (1991) Culture in language learning: A view from the United States. In De Bot, K., Ginsberg, R., & Kramsch, C. (éds). Foreign language research in cross-cultural perspective. John Benjamins

Kramsch, C. (1996) The Cultural Component of Language Teaching. In Zeitschrift für Interkulturellen Fremdsprachenunterricht, [En ligne], 1(2) Dernière Consultation juin 2014. L'URL : http://www.spz.tu-darmstadt.de/projekt_ejournal/jg_01_2/beitrag/kramsch2.htm

Kübler, N (1995) L'automatisation de la correction d'erreurs syntaxiques : application aux verbes de transfert en anglais pour francophones. Thèse de doctorat, Université Paris 7

Kübler, N & Cornu, E. (1994) Using automata to detect and correct errors in the written English of French-speakers. *TRANEL. Travaux Neuchâtelois de Linguistique*, 21, pp. 235-246

Kuteeva, M. & Mauranen, A. (2014) Writing for publication in multilingual contexts: An introduction to the special issue. Journal of English for Academic Purposes, 13, pp. 1-4

Lambrecht, K. (1994) Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents. Cambridge : Cambridge University Press

Lado, R. (1957). Linguistics across cultures: Applied linguistics for language teachers. Ann Arbor : University of Michigan Press

Lancaster, Zak. (2011). Interpersonal stance in L1 and L2 students' argumentative writing in economics: Implications for faculty development in WAC/WID programs. Across the Disciplines, 8 (4). Dernière consultation aout 2015, l'URL : http://wac.colostate.edu/atd/ell/lancaster.cfm

Lancaster, Z. (C. I.) (2012). Stance and Reader-Positioning in Upper-level Student Writing in Political Theory and Economics. Thèse de doctorat, Université de Michigan.

Lancaster, Z., & Olinger, A., R. (2014). Teaching Grammar in Context in College Writing Instruction: An Update on the Research Literature, WPA-CompPile Research Bibliographies, No. 24. WPA-CompPile Research Bibliographies Dernière consultation aout 2015, à l'URL http://comppile.org/wpa/bibliographies/Bib24/Grammar_in_Context.pdf.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. Biometrics. Vol. 33, pp. 159–174

Landragin, F., Poibeau, T. & Victorri, B. (2012) Analec : A New Tool for the Dynamic Annotation of Textual Data. Eighth International Conference on Language Resources and Evaluation, Istanbul, Turquie, 2012, pp. 357-362

Laufer, B (1994) Appropriation du vocabulaire : mots faciles, mots difficiles, mots impossibles. Acquisition et interaction en langue étrangère, 3 | URL : http://aile.revues.org/4895

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010) Automated Grammatical Error Detection for Language Learners. Morgan & Claypool Publishers

Leech, G. (1997). Introducing corpus annotation, In Garside, R., Leech, G., & McEnery, T. (éds). Corpus annotation: Linguistic information from computer text corpora, pp. 1-18. Londres & New York : Routledge

Leech, G. (2006). New resources, or just better old ones? The Holy Grail of representativeness. Language and Computers, 59 (1), pp. 133–149

Leech, G. (2011) Frequency, corpora and language learning. In Meunier F., De Cock S., Gilquin G. & Paquot M. (éds). A Taste for Corpora. In honour of Sylviane Granger, pp. 7-32. Amsterdam & Philadelphia : John Benjamins

Léon, J (2005). Claimed and Unclaimed Sources of Corpus Linguistics. Henry Sweet Society Bulletin, 2005, vol. 44, pp. 36-50

Li, C.N. & Thompson, S. (1976) Subject and Topic: New Typology of Language. In Charles N. Li (ed.). Subject and Topic. Londres & New York: Academic Press, pp. 457-61.

Li, P., Sepanski, S., & Zhao, X. (2006) Language history questionnaire: A web-based interface for bilingual research, Behavior Research Methods, 38 (2), pp. 202-210

Lin, Q. Y., Fawcett, R., & Davies, B. (1993) GENEDIS: the Discourse Generator in COMMUNAL. In Sloman, A., Hogg, D., Humphreys, G., Ramsay, A., & Partridge, D., (éds). Prospects for Artificial Intelligence: Proceedings of AISB 93. The Ninth Biennial Conference of the Society for the Study of Artificial Intelligence and the Simulation of behaviour. Amsterdam : IOS Press, pp. 148-57.

Long, M. L. (2005) Second Language Needs Analysis. Cambridge : Cambridge University Press

Luste-Chaa, O. (2009) Les acquisitions lexicales en français langue seconde : conceptions et applications. Thèse de doctorat, UniversitéPaul Verlaine-Metz

Lutkus, A. (1987) Problems in Measuring Syntactic Development: T-Units vs. Sentence Weight. Journal of Teaching Writing, vol.6 (1) pp. 49-67

Mahboob, A. & Knight, N. (éds). (2010) Appliable Linguistics. Continuum

Maingueneau, D. (1998) Analyser les textes de la communication. Paris : Dunod

Malrieu, D (2004) Linguistique de corpus, genres textuels, temps et personnes. Langages n° 153, pp. 73-85, URL : www.cairn.info/revue-langages-2004-1-page-73.htm.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M (2007) The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. Journal of Speech, Language, and Hearing Research, vol. 50, pp. 940–967

Marquilló Larruy, M. (2003) L'interprétation de l'erreur. Coll. Didactiques des langues étrangères. Paris : CLE International

Martin, J.R. (2003) Cohesion and Texture. In The Handbook of Discourse Analysis (éds). Schiffrin, D., Tannen, D & Hamilton, H. Blackwell Reference Online

Martin, J.R & Rose, D (2005) Designing literacy pedagogy: scaffolding democracy in the classroom. In Webster, J., Matthiessen, C. & Hasan, R. (éds). Continuing Discourse on Language. Londres: Continuum, pp. 251-280.

Martin, R. (2002) Comprendre la linguistique. Paris : Presses Universitaires de France, coll. « Quadrige ».

Matthiessen, C.M.I.M. (1983) Systemic grammar in computation: the Nigel case. In Proceedings of the First Annual Conference of the European Chapter of the Association for Computational Linguistics, pp. 155-164

Matthiessen, C.M.I.M. (2007) The Architecture of Language According to Systemic Functional Theory: Developments since the 1970s. In Hasan, R., Matthiessen, C. & Webster, J (éds). Continuing Discourse on Language: A Functional Perspective, vol. 2. Londres & Oakwille : Equinox, pp. 506-561

Matthiessen, C.M.I.M. & Bateman, J. A. (1991) Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese. Francis Pinter Publishers

Mauranen, A. (1993a) Cultural Differences in Academic Discourse – Problems of a Linguistic and Cultural Minority. In Lofman, L., Kurki-Suonio, L., Pellinen, S., & Lehtonen, J. (éds). Publications de l'association finlandaise de linguistique appliqué 51, pp. 157-174.

Mauranen, A. (1993b) Contrastive ESP Rhetoric: Metatext in Finnish-English Economics Texts. English for Specific Purposes, vol. 12, pp. 3-22.

Mauranen, A (1996) Discourse Competence – Evidence from Thematic Development in Native and Non-native Texts. In Ventola, E. & Mauranen, A. (éds). Academic writing: intercultural and textual issues. Amsterdam : John Benjamins.

Mauranen, A. (2006) A Rich Domain of ELF: The ELFA Corpus of Academic Discourse. Nordic Journal of English Studies 5 (2), pp. 145-59.

Mauranen, A. (2012) Exploring ELF: academic English shaped by non-native speakers. Cambridge : Cambridge University Press.

McEnery, T., & Wilson, A. (2001) Corpus linguistics (2ème édition). Édimbourg : Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006) Corpus-Based Language Studies, An Advance Resource Book. Londres & New York : Routledge

McEnery, T., & Hardie, A. (2012) Corpus linguistics: Method, theory and practice. Cambridge : Cambridge University Press

Miller, J (2005) Most of ESL Students have trouble with the articles. International Education Journal, 5(5), pp. 80-88

Moirand, S. (1979) Situation d'écrit. Paris : Clé International

Morley, D. (2000) Syntax in Functional Grammar: An Introduction to Lexicogrammar in systemic linguistics. Londres & New York : Continuum

Myhill, D., Jones, S., & Watson, A. (2013) Grammar matters: How teacher's grammatical knowledge impacts on the teaching of writing. Teaching and Teacher Education, 36, pp. 77-91

Nekula, M. (1991) Vilém Mathesius. In Verschueren, J., Östman, J.-O., Blommaert, J. & Bulcaen, Ch. (éds). Handbook of Pragmatics. Amsterdam & Philadelphia : John Benjamins, pp. 1–14.

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In Archer, D., Rauson, P., Wilson, A., & Mcenery, T. (éds). Actes du colloque 'Corpus Linguistics 2003'. UCREL 'technical paper'16, pp. 572-581.

O'Donnell, M. (2008) Demonstration of the UAM CorpusTool for text and image annotation. Proceedings of the ACL-08: HLT Demo Session (Companion Volume), Columbus, Ohio, juin 2008. Association for Computational Linguistics, pp. 13-16.

O'Donnell, M. (2010) UAM CorpusTool, version 2.8., disponible à l'URL http://www.wagsoft.com/CorpusTool/index.html

O'Donnell, M. & Bateman, J. (2005) SFL in Computational Contexts: a Contemporary History. In Webster, J. & Hasan, R., & Matthiessen, C. (éds). Continuing Discourse on Language: A Functional Perspective. Londres : Equinox, pp. 343-382.

O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., Stuart, K., & Boquera, M. (2009). Exploring the proficiency of English learners: The TREACLE project. Proceedings of the Fifth Corpus Linguistics, Liverpool

Osborne, J (2008) Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. In Granger, S. & Meunier, F. (éds). Phraseology in Foreign Language Learning and Teaching, Amsterdam & Philadelphia : John Benjamins, pp. 67–83.

Palmer, H. E (1917) The Scientific Study and Teaching of Languages. Londres : Oxford University Press

Palmer, H.E. (1921) The Principles of Language-Study. Yonkers-on-Hudson & New York : World Book Company

Palmer, H.E. (1924) Memorandum on Problems of English Teaching in the Light of a New Theory. Tokyo : Institute for Research in English Teaching

Palmer, H. E (1930) Interim Report on Vocabulary Selection: The Principles of Romanization. Tokyo : University of Tokyo

Palmer, H. E (1931) Second Interim Report on Vocabulary Selection. Tokyo : IRET

Palmer, H. E (1933) Second Interim Report on English Collocations. Japan : Kaitakusha

Perdue, C. (1980) L'analyse des erreurs : un bilan pratique. In Langages, 14, n° 57. pp. 87-94

Perry, E. (1993) Erreurs lexicales chez l'apprenant d'un niveau avancé, ASp, 2, pp. 81-91.

Pery-Woodley, M-P. (1993) Les écrits dans l'apprentissage, Paris : Hachette

Peytard, J & Moirand, S (1992) Discours et enseignement du français. Les lieux d'une rencontre. Collection Références, Paris : Hachette.

Pincemin, B. (2009) Panorama bref et pragmatique des outils de textométrie et apparentés, Fiche réalisée à l'intention des participants, Ecole thématique CNRS MISAT (Méthodes Informatiques et Statistiques en Analyse de Textes), Besançon, 15-19 juin 2009

Pincemin, B. (2011) Sémantique interprétative et textométrie – version abrégée, Corpus [en ligne], 10, URL : http://corpus.revues.org/2121, consulté le 24 août 2015

Piolat, A & Roussey, J-Y (1992) Rédaction de texte: éléments de psychologie cognitive. PsyCLE, Université de Provence

Polio, C. (1997) Measures of linguistic accuracy in second language writing research. Language Learning 47,1, pp. 101–143

Poudat, C. (2003) Outils de traitement de corpus. Texto! [en ligne], vol. VIII, n°2-3. URL : http://www.revue-texto.net/Corpus/Manufacture/pub/Poudat_Outils.html, consulté le 24 aout 2015

Reason, J. (1991) Human Error. Cambridge : Cambridge University Press

Reason, J. (2000) Human error: models and management. In Western Journal of Medecine, 172(6), pp. 393-396.

Reinertsen, J. L. (2000) Let's talk about error. In Western Journal of Medecine, 172 (6) pp. 356-357

Richards, J. C. (1974) (éds). Error Analysis: Perspectives on Second Language Acquisition. Londres : Longman.

Roehr, K. (2006) Metalinguistics knowledge and language-analytic ability in university-level L2 learners. In Essex Research Reports in Linguistics, 51, pp. 41-71

Rose, D. (1999) Culture, competence and schooling: approaches to literacy teaching in Indigenous school education. In F. Christie (éds). Pedagogy and the Shaping of Consciousness: Linguistic and Social Processes Londres : Cassell, pp. 217-245

Rück, H. (1991) Linguistique Textuelle Et Enseignement Du Français. Didier, Collection Langues & Apprentissages Langue.

Sanders, T. (2006) Text and text analysis. Utrecht University, Elsevier Ltd.

Sanders, T & Pander Maat, H (2006) Cohesion and Coherence: Linguistics approaches, Utrecht University, Elsevier ltd.

Sapir, E. (1929) The Status of Linguistics as Science. Ré-imprimé dans Mandelbaum, D (éds). (1949) Edward Sapir : Culture, Language, and Personality. Berkerley : University of California Press, pp. 65-77

Saussure, F. de (1979) Cours de linguistique générale, éd. originale : 1916. Paris : Payot

Schaeffer-Lacroix, E (2009) Corpus numériques et production écrite en langue étrangère. Une recherche avec des apprenants d'allemand. Thèse de doctorat, Université Sorbonne Nouvelle

Schiffrin, D., Tannen, D & Hamilton, H. (éds). (2001) The Handbook of Discourse Analysis. Oxford, Blackwell

Selinker, L. (1972). Interlanguage. In Jack C. Richards (éds). Error Analysis : Perspectives on Second Language Acquisition. Londres : Longman

Siepmann, D (2006) Academic Writing and Culture: An Overview of Differences between English, French and German. Meta: Translators' Journal, vol. 51(1), pp. 131-150

Sinclair, J. (2005). Corpus and Text - Basic Principles. In Wynne, M. (éds). Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books: pp. 1-16 [En ligne] à l'URL http://ahds.ac.uk/linguistic-corpora/ [Dernière consultation juillet 2014].

Smith, R. C. (1999) The Writing of Harold E. Palmer: An Overview. Hon-no-Tomosha Publishers. Tokyo : Hon-no-Tomosha

Smith, R. C. (2011) Harold E. Palmer's alternative 'applied linguistics'. Histoire Epistémologie Langage, vol.33 (1), pp. 53-67

Swales, J. (1990) Genre Analysis: English in academic and research settings. Cambridge : Cambridge University Press

Swales, J. (2009). Worlds of genre—metaphors of genre. In Bazerman, C., Bonini A., & Figueiredo D., (éds). Genre in a changing world, Perspectives on Writing pp. 147-157. Colorado : The WAC Clearinghouse and Parlor Press

Swales, J. (2011) Coda: Reflection on the future of genre and L2 writing, Journal of Second Language Writing, vol. 20 (1), pp. 83-85

Taguchi, N (2008). The Role Of Learning Environment In The Development Of Pragmatic Comprehension. Studies in Second Language Acquisition, 30, pp. 423-452

Taverniers, M. (2011) The syntax–semantics interface in Systemic Functional Grammar: Halliday's interpretation of the Hjelmslevian model of stratification. *Journal of Pragmatics* 43(4), pp. 1100–1126.

Taylor, J. R. (1997). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books.

Teich, E. (1999) Systemic Functional Grammars in Natural Language Generation. Linguistic Description and Computational Representation. Londres : Cassel

Thewissen, J. (2008) The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. In Proceedings from the 8th Teaching and Language Corpora Conference (TaLC 8), pp. 300-306.

Thwaite, A. (2015) Pre-service teachers linking their metalinguistic knowledge to their practice: a functional approach. Functional Linguistics 2015, 2 (4)

Torrance, M. (2006) Writing and cognition. In Brown, K. (éds). Encyclopedia of language and linguistics.. Oxford : Elsevier, pp. 679-682

Tremblay, M-C. (2006) Cross-Linguistic Influence in Third Language Acquisition : The Role of L2 Proficiency and L2 Exposure. In Cahiers linguistiques d'Ottawa. CLO/OPL, vol. 34, pp. 109-119

Turton, N & Heaton, J.B (1996) Longman Dictionary of Common Errors. Pearson, Longman

Ushida, E. (2005) The Role of Students' Attitudes and Motivation in Second Language Learning in Online Language Courses. CALICO Journal, 23 (1), pp. 49-78.

Van de Craats, I. (2002). The Role of the Mother Tongue in Second Language Learning. Babylonia, 10 (4), pp. 19-22.

VanDyke, J & Lehman, J. L (1997) An Architectural Account of Errors in Foreign Language Learning. In Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society.

Vasquez, M. C. A., (2005) Avoidance As A Learning Strategy. Estudios Ingleses de la Universidad Complutense, vol. 13 pp. 67-83

Ventola, E. & Mauranen, A (1991) Non-native writing and native revising of scientific articles. In Ventola, E. (éds). Functional and Systemic Linguistics : Approaches and Uses. Berlin, New York : Mouton de Gruyter, pp. 457-492

Webster, J. (2013) Appliable linguistic theory: Editor's Introduction. In Webster, J. (éds). About Halliday in the 21st century. Londres, New York : Bloomsbury

Weingart, S. N, Wilson, R. M., Gibberd, R. W., Harrison, B. (2000) Epidemiology of medical error. In Western Journal of Medicine 172(6): pp. 390–393

West, R (1994) Needs analysis in language teaching. Language Teaching, vol. 27(1), pp. 1-19

Wharton, S. (2012) Epistemological and interpersonal stance in a data description task : findings from a discipline-specific learner corpus. English for Specific Purposes, vol. 31(4), pp. 261-270.

Williams, G. (2006). La linguistique et le corpus: Une affaire prépositionnelle. Texto, revue de linguistique en ligne. Consulté juillet 2014 http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf

Wilson, G. M. (1909) Errors in Language of Grade Pupils, Educator-Journal

Wilson, G. M. (1920) Location the Language Errors in Children, The Elementary School Journal, 21, pp. 290-296.

Wilson, G. M. (1922) Language Error Tests, Journal of Educational Psychology, 13, pp. 430-37

Wulff, S., Römer, U., & Swales, J (2012) Attended/unattended this in academic student writing: Quantitative and qualitative perspectives. Corpus Linguistics and Linguistic Theory, vol. 8, 1, pp. 129–157

Wynne, M. (éds). Developing Linguistic Corpora: a Guide to Good Practice. Oxford : Oxbow Books: pp. 1-16 [En ligne] à l'URL http://ahds.ac.uk/linguistic-corpora/ [Dernière consultation juillet 2014]

ANNEXES

Annexe A1 : Questionnaire

Ce questionnaire fait partie d'une étude menée dans le cadre de mon doctorat en linguistique au laboratoire LaTTiCe. Nous demandons l'autorisation de consulter vos copies d'examen d'anglais de première année. NB : Aucune information personnelle ne sera retenue pendant ou après l'étude, et si vous autorisez la libre consultation de vos copies aujourd'hui, vous pourrez revenir sur votre décision ultérieurement.

1) Age :
2) Sexe :
3) Etes-vous né *ou* avez-vous passé toute votre vie en France ? Non Oui *[Entourez la bonne réponse]*
a) Si *non*, précisez les suivants :
i) Votre pays de naissance : _____
ii) Date d'arrivée en France (année d'installation) : _____
iii) Le(s) pays où vous avez vécu et la durée :

Pays	Durée

4) Quelle est la langue de votre pays de naissance ? :_____
5) Est-ce que cette langue est utilisée au foyer familial ? Non Oui *[Entourez la bonne réponse]*
a) Si *non*, précisez la ou les langues utilisée(s) : _____
b) Si *oui*, est-ce que c'est la seule langue utilisée ? : Non Oui *[Entourez la bonne réponse]*
i) Si vous avez répondez *non* à ***(5.b)*** précisez les langues employées : _____

6) Parlez-vous plus d'*une* langue ? Non Oui *[Entourez la bonne réponse]*
a) Si *oui*, pouvez-vous les énumérer en les classant, ainsi que les activités ci-dessous, selon l'échelle suivante :

Très rudimentaire	Rudimentaire	Moyen	Assez Bon	Bon	Très bon	Quasi-natif
1	2	3	4	5	6	7

Langues	Niveau global	Compréhension écrite (internet, livres)	Compréhension orale (radio, télé)	Expression écrite	Expression orale

b) Si vous écrivez dans une langue autre que le français, veuillez la préciser et la nature de vos écrits: _____
_____

7) Pouvez-vous préciser la ou les langue(s) principale(s) utilisée(s) pendant la période de scolarisation.

Niveau	**Langue(s) d'instruction**	**établissement classique, bilingue ou international**

		(précisez, s'il vous plaît)
Maternelle		
Primaire		
Collège		
Lycée		

Pour des cas spécifiques, précisez :

_____

_____

8) Quelle(s) langue(s) étrangères avez-vous étudiées au lycée ?

En LV1	
En LV2	
En LV3	
En option « renforcé »	
Autre(s) :	

9) Est-ce que vous poursuivez l'apprentissage de ces langues actuellement en dehors des cours obligatoires à l'université ? Non Oui. *[Entourez la bonne réponse]*

Si *oui*, précisez :

_____

_____

10) De quand remonte votre premier contact avec la langue anglaise ?

A la maternelle	
Au primaire	
Au collège	
Au lycée	
Autre, précisez :	

11) Avez-vous déjà étudié l'anglais en dehors des cours proposés au collège ou au lycée ? Non Oui

a) si *oui*, est-ce qu'il s'agissait des cours de :

Soutiens scolaires // cours particuliers	
Stages // formations intensives	
Préparation à un examen d'anglais (ex : PET, FCE, TOIEC, TOEFL ou autre)	
Autre, précisez	

b) Si autre, précisez : _____

_____

12) Avez-vous déjà suivi des cours enseignés entièrement en anglais ? Non Oui

(*Par exemple des cours d'économie en anglais*)

Si *oui*, précisez : _____

_____

13) En terme de fréquence, lesquelles des activités suivantes étiez-vous amené(e) à faire en cours d'anglais avant d'arriver à l'université ? *[Une réponse par activité]*

Activités	**1** jamais	**2** rarement	**3** souvent	**4** assez souvent	**5** très souvent
des exercices à trous					
des exercices de grammaire					
des réponses courtes (1 à 2 phrases)					
des traductions					
des textes courts (un paragraphe)					

des textes longs (plus d'un paragraphe)					

14) Avez-vous déjà séjourné dans un pays anglophone ? Non Oui

Si *oui*, est-ce qu'en raison :

Motif	**Où**	**Durée**
D'un séjour de vacances ?		
D'un séjour linguistique ?		
D'une formation (en anglais) ?		
Autre, précisez ici :		

15) Si vous avez des remarques que vous jugez utiles à faire sur votre parcours langagier, vous pouvez les ajouter ici :

_____

_____

Fait à _____

Le _____

Votre nom et prénom _____

Nous vous remercions d'avoir participé à l'enquête.

Clive Hamilton
Laboratoire LaTTiCe
CNRS : UMR 8094 - ENS Ulm - Paris III

Annexe A2 : Résultats du questionnaire

Cet abrégé présente les principaux résultats obtenus lors de la première étape de notre étude : à savoir l'étape des questionnaires. Notons que ces résultats sont d'une importance capitale puisqu'ils apportent la contextualisation nécessaire pour l'interprétation des erreurs relevées, notamment, dans les chapitres V, VI et VII. Il est tout d'abord question ici de dresser le portait linguistique de l'ensemble des sujets-participants retenus pour la présente étude. Pour ce faire, les résultats sont présentés de manière chronologique, et ce pour deux raisons. En effet, il s'agit de (i) faire un rappel pratique des nombreuses variables dépendantes et indépendantes sur lesquelles la présente étude s'est fondée et (ii) de montrer comment chaque partie est intrinsèquement interdépendante, facilitant de ce fait la compréhension de l'analyse qui sera développée par la suite. Autrement dit, les principaux points de différence relevés entre l'historique linguistique des sujets-participants (dans ce chapitre) seront nécessaires pour établir toute corrélation avec la systématicité de certaines erreurs observées dans les chapitres V et VI.

A.2 Les résultats obtenus du questionnaire

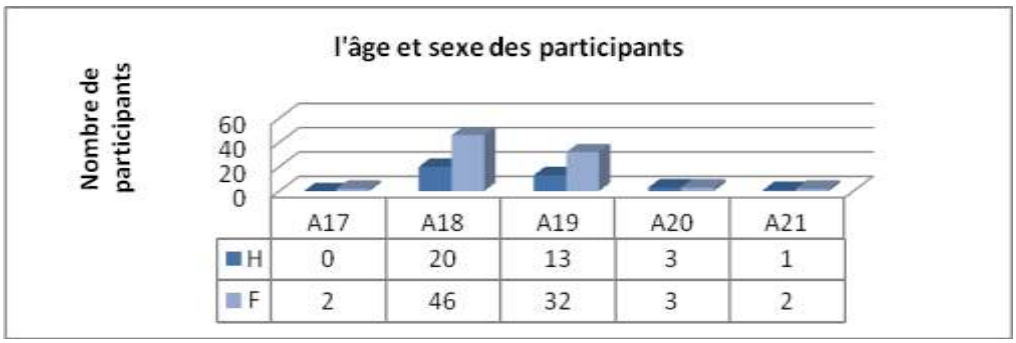
A2.1 Lés résultats obtenus du questionnaire

Dans un souci de brièveté, nous renvoyons aux sections 4.2.2.3 et 4.2.2.4 dans lesquelles la méthodologie du recueil et le tri des primo-répondants au questionnaire ont été préalablement décrits. A titre d'information, nous rappelons que les 164 répondants initiaux ont été réduits à 122 pour des raisons méthodologiques. De ce fait, l'ensemble des résultats présentés ci-dessous renvoie aux 122 sujets-participants retenus. Un exemple complet du questionnaire est fourni en annexe (cf. annexe A1 précédent).

A2.1.1 Les précisions sociodémographiques

Les questions allant de 1 à 6 répondent au besoin de recueillir des renseignements à caractère sociodémographiques et sociolinguistiques. Ces renseignements, rappelons-le, font partie des éléments ou paramètres que l'on souhaite croiser avec les résultats obtenus dans nos deux volets d'annotation (cf. section 4.2.4). Ce croisement cherche, en effet, à identifier toute corrélation entre

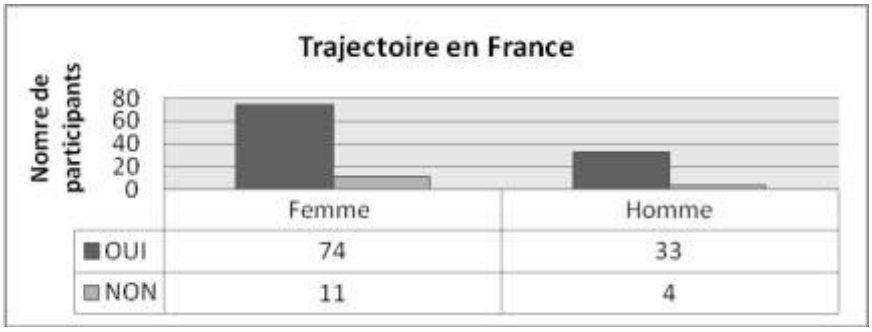
les facteurs sociodémographiques et sociolinguistiques et les différentes typologies d'erreurs qui seront présentées dans les chapitres V et VI. De ce fait, les 6 premières questions portent sur l'âge, le sexe, le pays de naissance ou de scolarité des sujets-participants, sans oublier la ou les langue(s) employée(s) au foyer. La figure 41 ci-après illustre la répartition de l'âge selon le sexe des participants. Parmi les 122 participants, 85 sont de sexe féminin et 37 sont de sexe masculin, correspondant respectivement à 70% et 30%. La tranche d'âge observée est de 17 à 21 ans, avec une moyenne qui se situe à 18,5 ans aussi bien pour les hommes que pour les femmes.



	A17	A18	A19	A20	A21
H	0	20	13	3	1
F	2	46	32	3	2

Figure 41: la distribution de l'âge et le sexe des participants

Nous avons demandé aux répondants de préciser s'ils sont nés ou ont passé toute leur vie en France. Pour rappel, cette question est une des premières visant à établir l'historique langagier des sujets-participants et renvoie singulièrement à la notion de contact de langues : une des variables dépendantes de notre étude. Le but étant d'établir un profil précis des différents contacts de langues et leur incidence sur l'apprentissage de l'anglais et également d'identifier toute corrélation possible entre le type d'erreurs commises et le parcours linguistique des participants. Le graphique ci-après renvoie donc aux réponses fournies à cette question. Le pourcentage de ceux ayant répondu par l'affirmative s'élève à 88% et demeure proportionnel si l'on sépare les répondants selon leur sexe.



	Femme	Homme
OUI	74	33
NON	11	4

Figure 42 : La répartition des participants nés ou ayant grandi en France

Parmi les 15 participants ayant signalé ne pas être nés ou ne pas avoir passé toute leur vie en France, le tableau suivant dresse l'état des réponses en apportant des précisions quant aux pays et la durée de leur séjour. C'est ainsi que nous relevons le fait que les profils ne sont pas homogènes. En

effet, au vu de ces premiers éléments d'information, il est tout à fait légitime de se demander si ces différences de trajectoire pourraient avoir une incidence réelle sur l'output final, à savoir la production concrète en anglais langue étrangère. De plus, si l'on pousse ce raisonnement un peu plus loin, on pourrait également chercher à savoir si ces trajectoires permettent la polarisation de certains types d'erreurs (ou à l'inverse, l'absence de certains types d'erreurs) chez d'autres sujets-participants dont la trajectoire linguistique s'avère bien plus linéaire.

pays où ils ont vécu	**durée en année**
Indonésie	4
Grèce	17,5
Maroc	17
Maroc	18
Maroc	18
FR ; Italie ; FR ; E.-U ; FR	1 ; 6 ; 7 ; 1 ; 3
RU ; Ecosse	1 ; *8 mois
Ukraine : Allemagne ; FR	8 ; 3 ; 8
Tahiti	14
Indonésie ; Russie ; FR	5 ; 1 ; 13
FR ; E.-U	20 ; 1
E.-U ; FR	2 ; 17
À l'étranger	???

Tableau 61 Trajectoire des sujets-participants ayant vécu en dehors de la France

Soulignons, dans le tableau 61, que sur les quinze répondants n'ayant pas passé toute leur vie en France, douze ont précisé leur pays de naissance et l'endroit où ils ont grandi, tandis que 3 ont tout simplement signalé être nés à l'étranger. Notons également, à titre accessoire, qu'un des postulats sous-jacents à notre étude est que les sujets-participants n'ayant pas été scolarisés en France ne commettront pas forcément les mêmes types d'erreurs et pourraient même maîtriser (ou pas) des points que la majorité des répondants n'auraient pas encore acquis – en raison de leur contact renforcé avec plusieurs langues[140]. Nous nous attendons donc à vérifier si ce postulat s'avère probant ou si peu de différences émergent de ces profils par rapport aux autres. Selon le tableau 62, il faut donc comprendre que 5 participants sont nés en France mais ont passé leur vie à l'étranger ; 3 sont nés et ont grandi au Maroc avant de s'inscrire dans l'enseignement supérieur en France, et ainsi de suite. Il est à noter toutefois que certains des répondants (au nombre de 5) ont indiqué être nés à l'étranger mais ont grandi ou passé toute leur vie en France : ils n'ont pas été comptabilisés

[140] Chez ces participants en particulier, il y a plusieurs langues qui entrent en contact : à savoir la langue du pays de naissance (qui est souvent la langue de leur « première » scolarisation), le français langue de scolarisation dans l'enseignement supérieur et l'anglais langue étrangère.

en tant que participants étrangers par la suite. C'est notamment le cas des participants nés au Luxembourg et au Mali, par exemple.

pays de naissance	***nb* de participants**
France	5
Grèce	1
Luxembourg	1
Mali-Bamako	1
Maroc	3
Ukraine	1
Non connu	3

Tableau 62 : le pays de naissance de sujets-participants n'ayant pas vécu en France

Par ailleurs, afin d'approfondir la notion de contact des langues chez nos sujets-participants, il a fallu obtenir des précisions sur les moyens linguistiques dont ils disposaient dans le cadre privé - c'est-à-dire dans leur vie personnelle par opposition au cadre institutionnalisé. Des questions ont donc été formulées en ce sens : à savoir « quelle est la langue de votre pays de naissance » ; « est-ce que cette langue est parlée à la maison » ; et « est-ce que vous en parlez d'autres à la maison ». Le tableau 63 ci-dessous illustre l'ensemble des réponses obtenues. La première colonne de gauche renvoie aux langues signalées comme étant la langue officielle du pays de naissance de nos 122 sujets-participants. Les trois autres colonnes établissent si la langue en question est parlée ou non à la maison.

	Langue parlée à la maison			
Langue du pays de naissance	**Non**	**Oui**	**Non-connu**	***total***
Anglais	1	0	0	*1*
Arabe	1	3	0	*4*
Grec	0	1	0	*1*
Français	4	104	1	*109*
Français + Bambara	0	1	0	*1*
Luxembourgeois	0	1	0	*1*
Ukrainien	0	1	0	*1*
Non-connu	0	0	4	*4*
total	***6***	***111***	***5***	***122***

Tableau 63 : Rapport de force entre langue du pays et langue à la maison (1ère partie)

Contrairement à ce qui avait été était escompté, ce tableau n'a pas permis de dégager ou de regrouper plusieurs profils bien distincts les uns des autres. Il en est néanmoins ressorti que six des répondants n'utilisent pas la langue de leur pays de naissance ; cinq n'ont pas fourni d'informations complètes - soit en indiquant la langue de leur pays de naissance sans préciser si celle-ci est utilisée

à la maison, soit en ne rien précisant du tout à savoir, ni la langue du pays ni celle(s) utilisée(s) à la maison. De plus 111 sur 122 sujets-participants affirment utiliser la langue du pays à la maison, parmi lesquels on trouve uniquement 104 répondants pour la langue française.

Pour les six participants ayant répondu que la langue du pays n'était pas utilisée à la maison, on a demandé à ces derniers d'indiquer la langue employée à la maison. Le tableau 64 fournit un aperçu de leurs réponses. Ce que l'on doit comprendre, ici, est que la langue du pays ne coïncide pas toujours avec la langue utilisée hors du cadre officiel ou institutionnalisé. Par exemple, certains des sujets-participants affirment être nés en France, mais n'utilisent pas le français à la maison.

Langues du pays	*N=**6***	*Langue utilisée à la maison*
Anglais	1	français
Arabe	1	français
Français	1	wenzhai (dialect chinois)
Français	1	tamoul
Français	1	italien
Non connu	1	Non connu (sp_68)

Tableau 64: Rapport de force entre langue du pays et langue à la maison (2ème partie)

De surcroit, 86 participants ont précisé n'utiliser qu'une seule langue à la maison tandis que 16 ont signalé utiliser au moins une autre en plus de la langue du pays de naissance. Nous rappelons ici que ces croisements d'informations nous permettent d'explorer toute corrélation entre les types de contacts linguistiques et la maîtrise de l'anglais – et plus particulièrement entre les différents types d'erreurs commises qui constituent le point de départ de notre analyse. Cette sous-population, au contact d'une autre langue de façon non-institutionnelle, pourrait alors être comparée à celle n'ayant que le français à la fois comme langue du pays de naissance et comme seule langue employée à la maison.

Langues du pays	***langues utilisées à la maison***
grec	français
ukrainien	français, ukrainien, russe
arabe	français
arabe	français
français	basque
français	néerlandais
français	hongrois
français	cambodgien
français	hindi
arabe	français
français	arabe
français	espagnol
français	créole mauricien, hindi

français	cambodgien
français	français, italien, créole
français	hébreu

Tableau 65: Répartition des différentes langues utilisées à la maison

A2.1.2 Les parcours linguistiques institutionnels

Pour approfondir la notion de contact des langues chez nos sujets-participants, nous avons cherché à faire un inventaire des langues répertoriées chez chacun d'entre eux. Il en ressort que parmi les 122 sondés, seul un répondant affirme ne parler qu'une seule langue : à savoir le français. Par contre, 8 déclarent parler une langue de plus que celle du pays de naissance ou celle utilisée à la maison ; le chiffre s'élève à 78 pour deux langues ; 25 pour 3 langues et 10 pour 4 langues complémentaires. La figure 43 ci-après permet de visualiser ces chiffres en pourcentage.

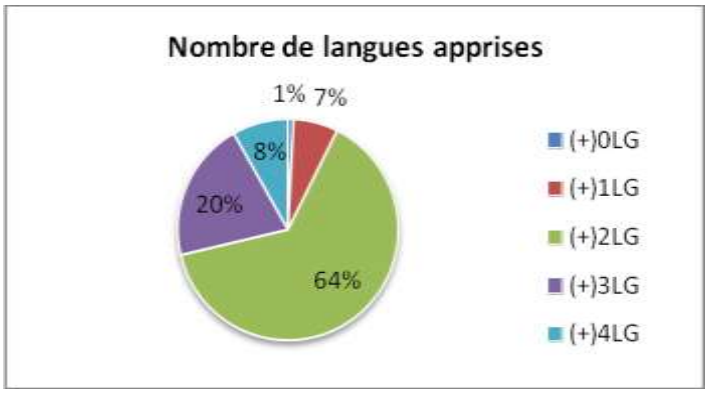


Figure 43: La répartition des langues parlées par les participants

L'anglais est la seule langue citée par 120 participants, comme étant une langue complémentaire à celle parlée à la maison. Une précision toutefois s'impose : aucun des 122 participants n'a indiqué parler anglais à la maison ou en dehors du cadre des séjours à l'étranger ou des formations institutionnalisées. Etant donné que ce résultat était en quelque sorte attendu ici, les répondants ont dû se soumettre à une grille d'auto-évaluation en anglais afin d'obtenir une analyse plus fine. La courbe ci-dessous met en avant les précisions obtenues.

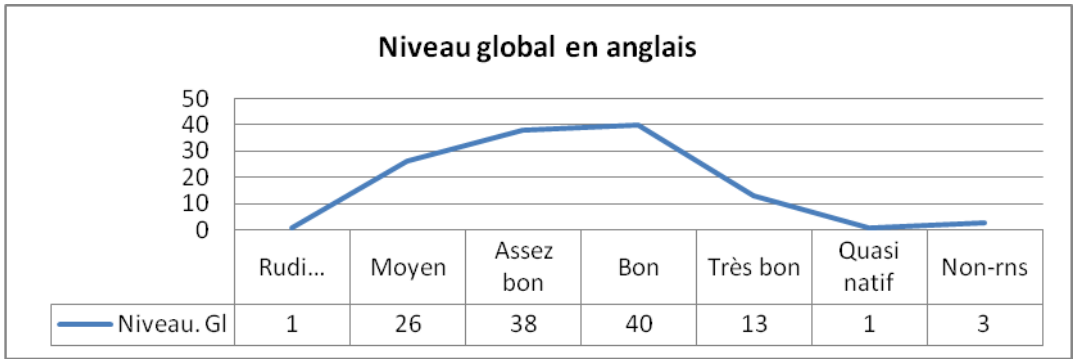


Figure 44 : La perception du niveau global en anglais des sujets-participants

La courbe montre le niveau global des répondants en anglais : allant d'un niveau dit rudimentaire à un niveau jugé quasi-natif. En effet, comme nous pouvons le voir presque deux répondants sur trois évaluent leur anglais comme étant 'assez bon' ou 'bon'. Alors que seulement 11% pensent avoir un niveau qu'ils qualifieront de 'très bon' ; voire 0,8% pour un niveau 'quasi-natif'. Les répondants ont également dû apporter des précisions sur leur « niveau global » en anglais, en indiquant comment ils évalueront leurs quatre compétences linguistiques : à savoir la compréhension orale, la compréhension écrite, l'expression orale et l'expression écrite.

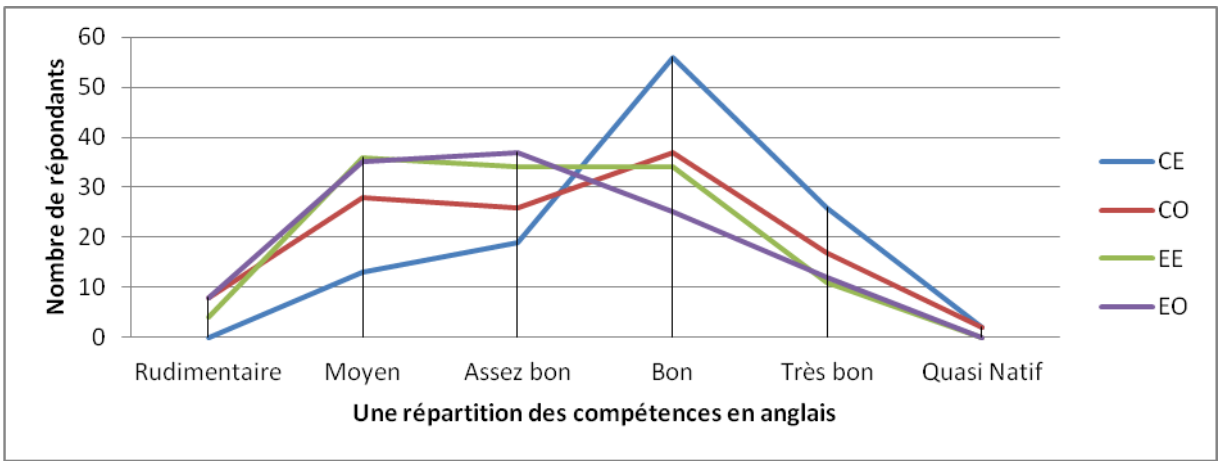


Figure 45 : La répartition des quatre compétences en anglais

Ce graphique met en exergue plusieurs types d'informations auxquels nous ne nous attendions pas, du moins pas de manière si tranchée. En effet, très peu de répondants considèrent qu'ils ont un 'très bon' niveau d'expression : 10% pensent avoir une très bonne expression orale (EO) et 9% une très bonne expression écrite (EE). Dans ces deux compétences, la majorité (à savoir 68% pour l'expression orale et 62% pour l'expression écrite) s'est attribué des notes allant de 1 à 3, sachant que la note 6 représente un niveau 'quasi-natif'. Ce résultat nous permet d'avoir un aperçu de l'opinion de ces répondants concernant leurs propres compétences. Notons aussi qu'en compréhension orale (CO) et compréhension écrite (CE), le taux s'élève respectivement à 52% et 11% pour ceux qui se sont attribué des notes de 1 à 3. Tandis que 72% des répondants se sont attribué des notes de 4 à 6 pour la compréhension écrite. Ce que nous retenons donc de cette auto-évaluation est que les sujets-participants pensent avoir des compétences actives en lecture mais reconnaissent avoir des difficultés dans les activités de production.

Il en est également ressorti du questionnaire que la langue d'instruction du parcours scolaire des sujets-participants était une variable à ne pas négliger. En effet, on relève un taux de scolarisation en langue française qui varie petit à petit, au fur et à mesure que l'on progresse dans le contexte institutionnel. C'est-à-dire, on recense 112 répondants ayant eu le français comme unique langue

d'instruction à la maternelle ; ce chiffre tombe à 99 en école primaire et 91 au collège pour finir à 88 au lycée. Autrement dit, il y a eu une diminution progressive, qui atteint même 21%, de ceux qui ont eu deux langues d'instruction formelle.

Toutefois, il est également important de noter qu'un seul répondant affirme avoir suivi une scolarisation à la fois en français et en anglais, et ce dès la maternelle. Ce chiffre s'élève à 23 au lycée. Cette tendance est plus nuancée chez les participants ayant commencé avec le français et une autre langue étrangère à la maternelle ou ceux qui n'avaient pas du tout le français comme langue d'instruction en début de leur parcours scolaire. Ces derniers sont représentés dans la figure 46 comme AUT+AUT.

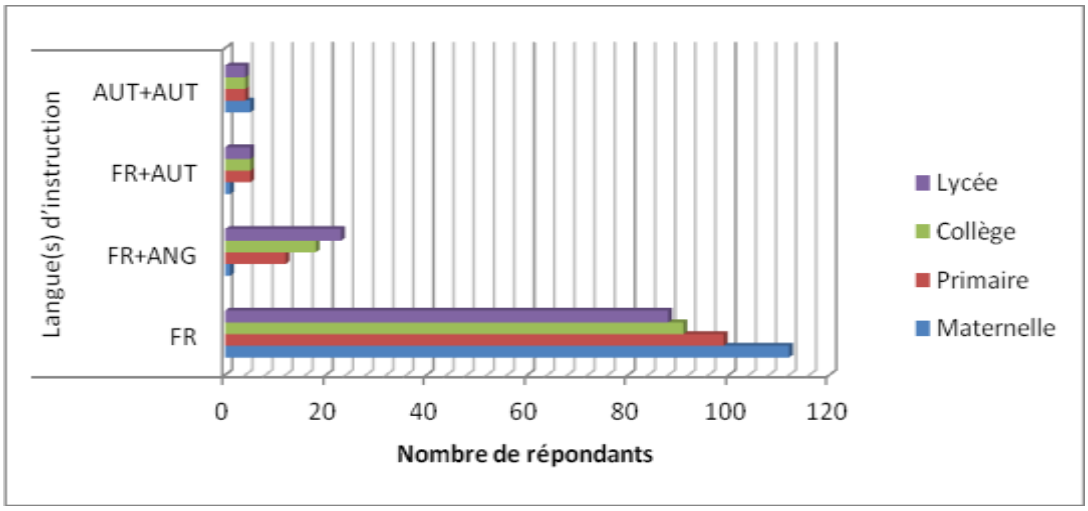


Figure 46 : Rapport de force entre langue d'instruction et parcours scolaire

Concernant l'initiation à la langue anglaise ou le premier contact dans un cadre institutionnel, presque trois quarts des participants affirment l'avoir fait en école primaire - notamment dans les deux dernières années, à savoir en cours moyen 1re année et cours moyen 2e année (communément abrégé CM1 ou CM2). Toutefois, 17 des 122 répondants soulignent avoir eu un premier contact à la maternelle ; 16 des 122 au collège ; et 1 seul répondant au lycée. Ce contact précoce, voire privilégié, peut expliquer en partie le fait qu'une grande majorité des sujets-participants choisissent l'anglais comme première langue vivante étrangère (abrégé LV1) au collège. Cette tendance se vérifie par le nombre élevé de personnes ayant choisi l'anglais en LV1 au lycée (109 sujets-participants) ; ils ne sont que quinze à l'avoir choisi en LV2 ; quatre en LV3 ; trois en 'autre option' n'entrant pas dans les cadres précédemment établis. Ils sont tout de même neuf à avoir choisi l'anglais parmi les trois premières options tout en y ajoutant l'‘option renforcée’, en complément de la formation obligatoire.

Cet engouement se matérialise de manière plus ostentatoire par le nombre ayant suivi des cours de langue anglaise complémentaire, c'est-à-dire en dehors du cadre institutionnel obligatoire au collège et au lycée. En effet, c'est notamment le cas de 50 % de la population d'étude. De plus, ils sont 18 à avoir suivi des cours particuliers ; 32 à avoir suivi des stages ou des formations intensives ; 23 à avoir suivi des ateliers de préparation pour des certifications en anglais - du type FCE, TOEIC, TOEFL[141], et ainsi de suite. Notons à titre d'information que 23 répondants affirment avoir cherché à peaufiner leurs connaissances en anglais par le biais de voyages linguistiques. Tout ceci démontre à quel point - à l'intérieur d'un même groupe - les écarts de parcours peuvent être variés. D'un côté, on pourrait presque « se féliciter » d'avoir rassemblé des profils aussi hétérogènes – ce qui correspond à notre besoin d'avoir un corpus représentatif (cf. section 4.1.1.1). Mais d'un autre côté, si l'on veut établir des « erreurs types » selon un parcours donné, ces nombreux écarts de trajectoire ne se prêtent pas commodément à ce type de comparaison, en raison des multiples dissimilitudes.

Le dernier point sur l'apprentissage institutionnalisé de l'anglais dans notre questionnaire réside dans le type de productions auquel les participants ont été habitués. Étant donné que notre corpus d'étude est un corpus écrit, nous nous intéressons, de ce fait, uniquement aux pratiques relevant de la production écrite. Il est question donc de s'interroger sur la fréquence des différents types d'exercices et les pratiques rédactionnelles mis en avant dans les cours d'anglais langue étrangère. Le but étant d'examiner toute incidence sur les différents types d'exercices auxquels nos sujets-participants sont habitués et les productions écrites recueillies dans le cadre de notre corpus d'étude.

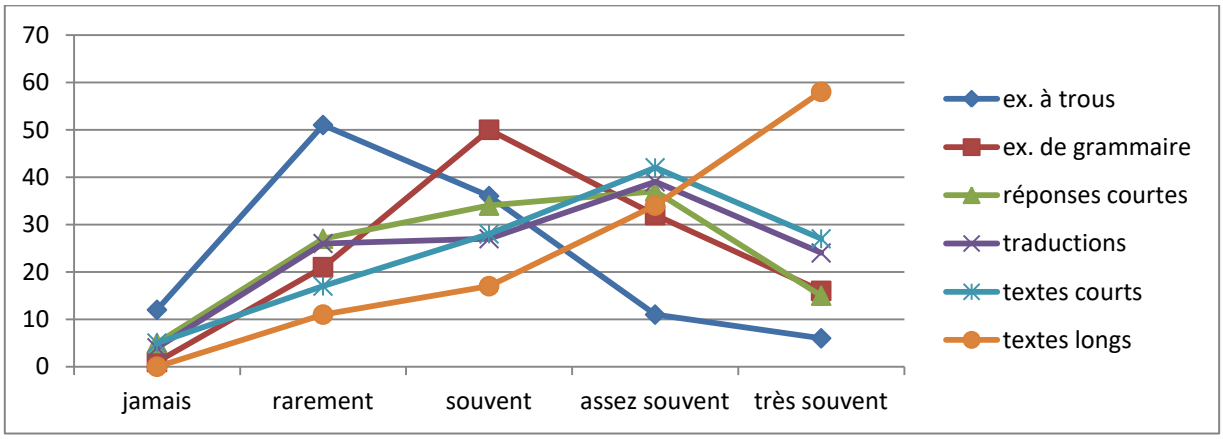


Figure 47 : La fréquence des différents types d'activités rencontrées en classe d'anglais langue étrangère

[141] A savoir, First Certificate in English (FCE) ; Test of English for International Communication (TOEIC) et Test of English as a Foreign Language (TOEFL)

En termes de fréquence, la figure 47 met en avant le fait que les exercices à trous sont assez rares dans les pratiques évaluatives. Les exercices de grammaire sont statistiquement très fréquents : deux tiers considèrent qu'ils ont ce type d'exercice de manière 'souvent' ou 'assez souvent'. Les réponses sont plus nuancées pour les exercices de type 'traductions' et de type 'réponses courtes' avec 54% et 58% des participants signalant respectivement leur fréquence comme 'souvent' ou 'assez souvent'. Ce qui retient notre attention, par contre, est le signalement de l'exercice de rédaction, à la fois en tant que textes courts et textes longs, ce qui revêt un caractère plus marqué ici : à savoir 56 % de notre population d'étude considère avoir affaire à ces types de rédaction 'assez souvent' voire 'très souvent' pour le premier et jusqu'à 75% portant le même jugement pour les rédactions longues. Cela étant, l'activité de rédaction ne peut donc pas être considérée comme un frein en soi, au vu de son caractère « fréquent » chez les participants. A ce titre, le processus rédactionnel n'est pas étudié comme un facteur digne d'intérêt dans la présente étude.

A2.1.3 Les contacts linguistiques ampliatifs

Comme nous l'avons souligné ci-dessus, il y a une sorte d'engouement patent dans l'apprentissage de l'anglais chez une grande majorité des sujets-participants. Ceci se manifeste dans le fait de chercher à consolider leur apprentissage quasi-obligatoire de cette langue par des compléments de formation : soit par le biais, par exemple, de cours de soutien scolaire, soit par des séjours plus au moins longs dans des pays anglophones. En effet, 97 des 122 répondants ont séjourné dans des pays anglophones contre 12 qui indiquent ne l'avoir jamais fait et 13 non-renseignés. Le tableau suivant recense la durée moyenne et la raison principale évoquée.

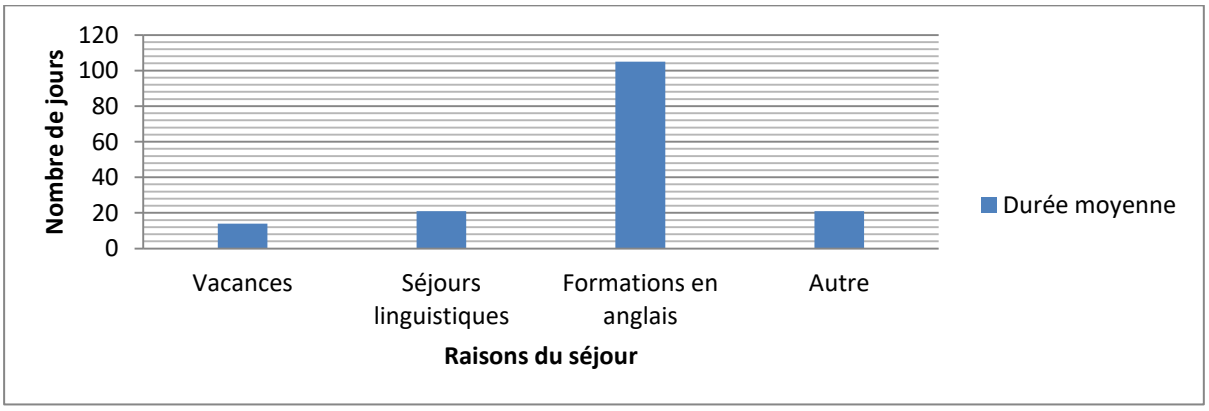


Figure 48 : Un croisement entre les raisons et la durée des séjours dans les pays anglophones

Toutefois, malgré une durée moyenne des séjours passés dans des pays anglophones évidente, des précisions complémentaires s'imposent. En effet, la durée individuelle varie entre moins de 10 jours et peut atteindre 3 mois pour des séjours effectués pour des raisons de vacances ; 10 jours est la valeur minimale pour les séjours linguistiques avec une valeur maximale allant jusqu'à 6 mois ;

la disparité est encore plus flagrante entre ceux qui ont effectué des séjours pour des raisons de formation en anglais – où le temps varie entre une semaine et une année entière et ceux ayant effectué des séjours pour des motifs divers – où la valeur temporelle varie à nouveau entre 10 jours et 2 mois.

Cela étant, pour chacune des quatre raisons invoquées, les participants situés aux deux extrémités de chaque groupe ou catégorie pourront être comparés les uns aux autres. Autrement dit, dans la catégorie 'séjour linguistique', il pourrait être question d'utiliser la variable temporelle comme variable indépendante pour mesurer l'écart des erreurs entre (i) un participant n'ayant pas effectué de séjour dans un pays anglophone ; (ii) un autre qui aurait effectué le plus court séjour (10 jours) ; et enfin, (iii) un autre ayant effectué le plus long séjour (6 mois). Ces trois participants seraient donc considérés comme des variables dépendantes.

A2.2 Le bilan du questionnaire

En définitive, les résultats obtenus grâce à ces 122 questionnaires ont permis de dresser un profil linguistique assez représentatif voire très précis sur l'ensemble des répondants. Cependant, les nombreuses variations notamment dans les parcours scolaires des uns et des autres ne permettent pas de regrouper la totalité des participants dans des sous-ensembles homogènes. Il est néanmoins possible d'en établir trois sous-groupes. Ces trois groupes sont définis ci-après :
[P1] Participants étrangers (tout type confondu) (E+1CLR)
[P2] Français sans contact linguistique régulier ou « important » (F-CLR)
[P3] Français avec contact linguistique régulier en dehors du cadre institutionnel (F+CLR)

- (E+1CLR) fait allusion aux participants étrangers ayant suivi la plus grande partie de leur scolarité à l'étranger. Par exemple le sujet-participant (sp_033) a vécu uniquement en Grèce avant de s'inscrire dans l'enseignement supérieur français.
- (F-CLR) désigne les participants francophones – qui sont nés ou ont passé toute leur vie en France – et qui n'ont ni suivi des cours entièrement en anglais ni séjourné dans un pays anglophone. Les participants retenus dans cette catégorie ne parlent que le français à la maison.
- (F+CLR) revoie à ceux ayant eu un parcours semblable au (P2 ou F-CLR) mais ayant séjourné dans des pays anglophones pour des durées supérieures à deux mois.

La comparaison de ces groupes vise à établir toute corrélation entre le type de contact, la fréquence d'un contact linguistique prolongé et les différentes typologies d'erreurs identifiées dans les chapitres V et VI.

Annexe A3 : Schéma d'annotation d'UAM CorpusTool, v.2.8. (Adapté)

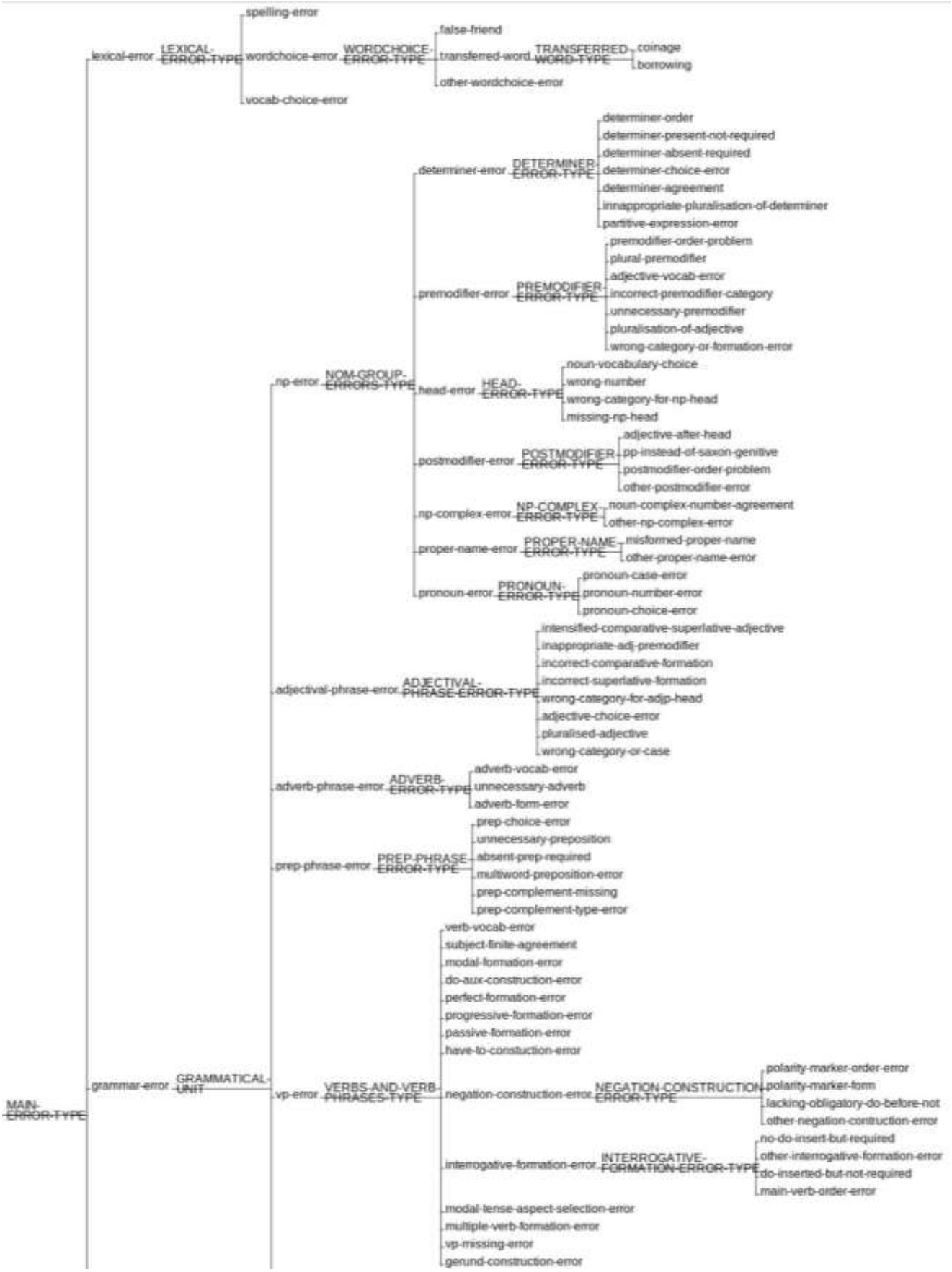


Figure 49 : Schéma d'annotation d'UAM en entier (1/2)

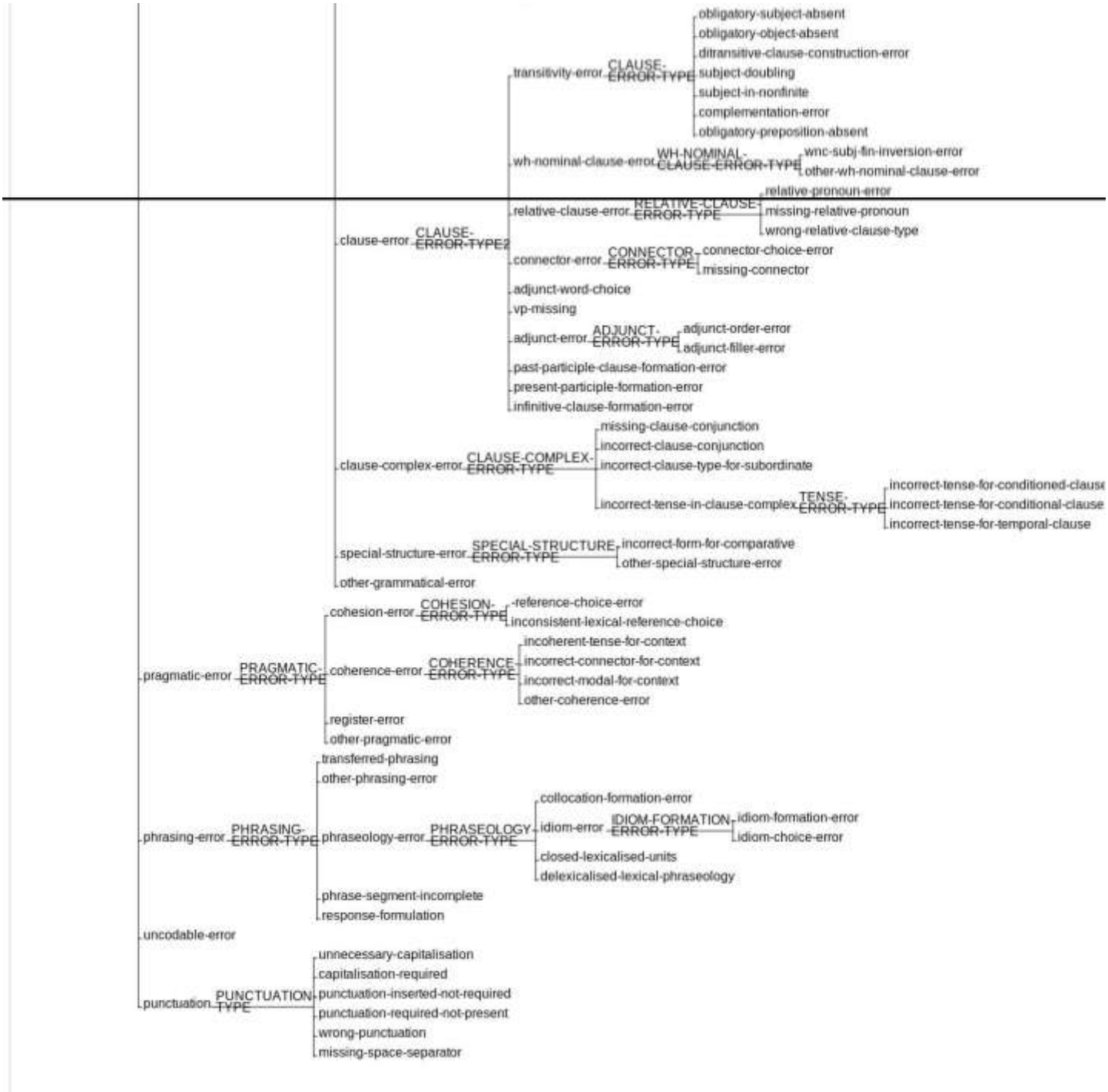
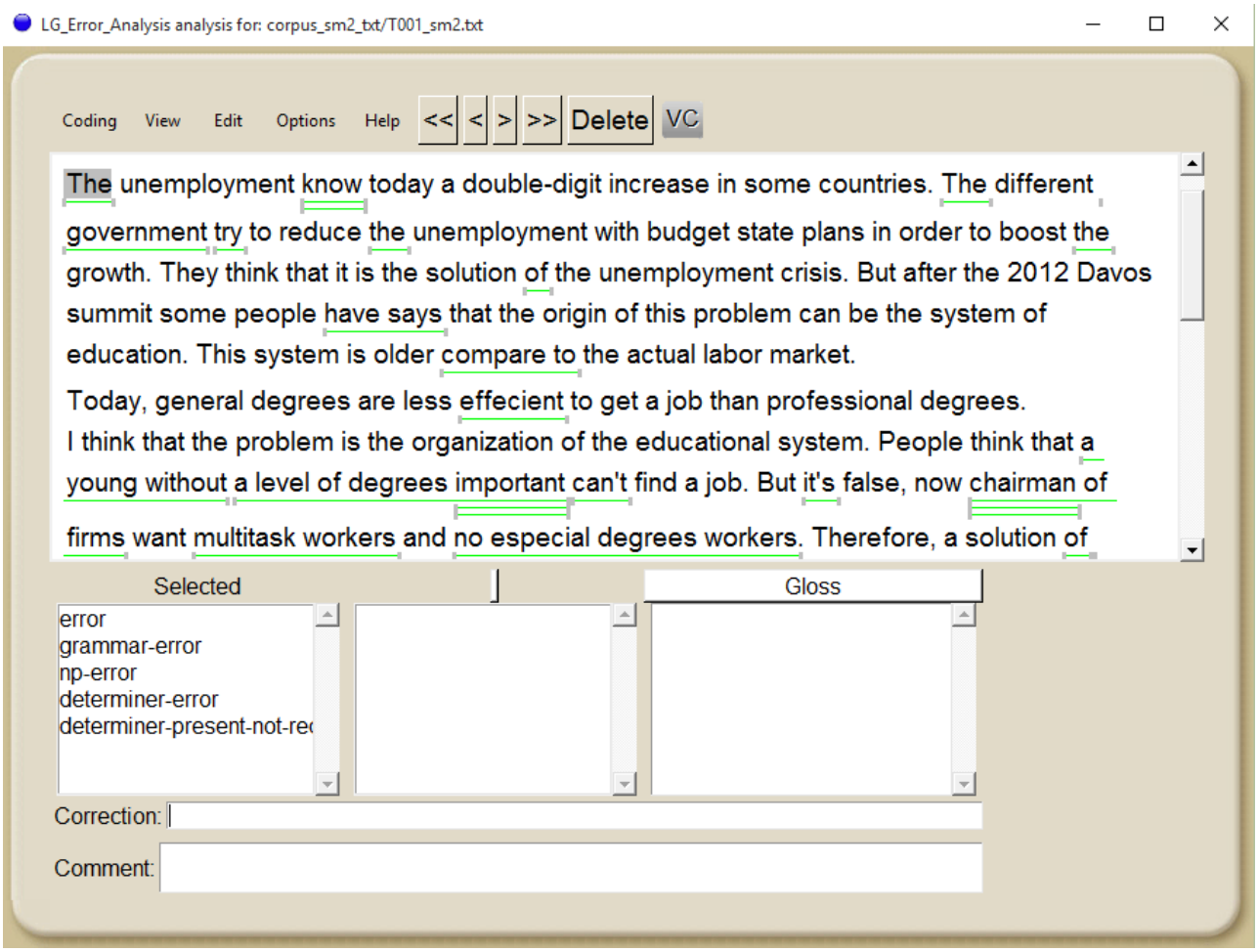


Figure 50 : Schéma d'annotation d'UAM en entier (2/2)

Annexe A4 : Aperçu d'un texte annoté



Annexe A5 : Exemple d'un fichier entier annoté en format XML

```
- <document>
  - <header>
      <textfile>corpus_sm2_txt/T001_sm2.txt</textfile>
      <lang>english</lang>
    </header>
  - <body>
      <segment id="1" features="error;grammar-error;np-error;determiner-error;determiner-present-not-required" state="active">The</segment>
      unemployment
    - <segment id="3" features="error;grammar-error;vp-error;subject-finite-agreement" state="active">
        <segment id="2" features="error;grammar-error;vp-error;verb-vocab-error" state="active" parent="3">know</segment>
      </segment>
      today a double-digit increase in some countries.
      <segment id="4" features="error;grammar-error;np-error;determiner-error;determiner-present-not-required" state="active">The</segment>
      different
      <segment id="9" features="error;grammar-error;np-error;head-error;wrong-number" state="active">government</segment>
      <segment id="5" features="error;grammar-error;vp-error;subject-finite-agreement" state="active">try</segment>
      to reduce
      <segment id="6" features="error;grammar-error;np-error;determiner-error;determiner-present-not-required" state="active">the</segment>
      unemployment with budget state plans in order to boost
      <segment id="7" features="error;grammar-error;np-error;determiner-error;determiner-present-not-required" state="active">the</segment>
      growth. They think that it is the solution
      <segment id="8" features="error;grammar-error;prep-phrase-error;prep-choice-error" state="active">of</segment>
      the unemployment crisis. But after the 2012 Davos summit some people
      <segment id="10" features="error;grammar-error;vp-error;perfect-formation-error" state="active">have says</segment>
      that the origin of this problem can be the system of education. This system is older
      <segment id="11" features="error;grammar-error;special-structure-error;incorrect-form-for-comparative" state="active">compare to</segment>
      the actual labor market. Today, general degrees are less
      <segment id="12" features="error;lexical-error;spelling-error" state="active">effecient</segment>
      to get a job than professional degrees. I think that the problem is the organization of the educational system. People think that
      <segment id="14" features="error;grammar-error;np-error;head-error;missing-np-head" state="active">a young without</segment>
    - <segment id="16" features="error;phrasing-error;phrase-segment-incomplete" state="active">
        a level of degrees
      - <segment id="15" features="error;grammar-error;adjectival-phrase-error;adjective-choice-error" state="active" parent="16">
          <segment id="13" features="error;grammar-error;np-error;postmodifier-error;adjective-after-head" state="active" parent="15">important</segment>
        </segment>
```

Figure 51 : Exemple d'un fichier annoté (1/3)

```
      </segment>
      <segment id="17" features="error;pragmatic-error;register-error" state="active">can't</segment>
      find a job. But
      <segment id="18" features="error;pragmatic-error;register-error" state="active">it's</segment>
      false, now
    - <segment id="21" features="error;grammar-error;np-error;postmodifier-error;pp-instead-of-saxon-genitive" state="active" comment="several different options possible">
      - <segment id="20" features="error;grammar-error;np-error;head-error;noun-vocabulary-choice" state="active" parent="21">
          <segment id="19" features="error;grammar-error;np-error;determiner-error;determiner-absent-required" state="active" parent="20">chairman</segment>
        </segment>
        of firms
      </segment>
      want
      <segment id="22" features="error" state="active">multitask workers</segment>
      and
    - <segment id="24" features="error;phrasing-error;phrase-segment-incomplete" state="active" comment="and not workers with professional or specialized degrees?">
        no
        <segment id="23" features="error;lexical-error;spelling-error" state="active" parent="24">especial</segment>
        degrees workers
      </segment>
      . Therefore, a solution
      <segment id="25" features="error;grammar-error;prep-phrase-error;prep-choice-error" state="active">of</segment>
      <segment id="26" features="error;grammar-error;np-error;determiner-error;determiner-absent-required" state="active">employment crisis</segment>
      can be a modification of education.
    - <segment id="29" features="error;phrasing-error;phrase-segment-incomplete" state="active">
        The
        <segment id="27" features="error;lexical-error;spelling-error" state="active" parent="29">begining</segment>
      </segment>
      is the promotion of professional degrees
      <segment id="31" features="error;grammar-error;adjectival-phrase-error;incorrect-comparative-formation" state="active">more</segment>
    - <segment id="32" features="error;grammar-error;adjectival-phrase-error;adjective-choice-error" state="active">
        <segment id="30" features="error;lexical-error;spelling-error" state="active" parent="32">specifics</segment>
      </segment>
      than general degrees.
      <segment id="33" features="error;pragmatic-error;cohesion-error;-reference-choice-error" state="active">This</segment>
      professional education
      <segment id="34" features="error;grammar-error;vp-error;subject-finite-agreement" state="active">have</segment>
```

Figure 52 : Exemple d'un fichier annoté (2/3)

have
a strong link with the labour world.
Students learn the different actions and advices
which
can be apply
in their
futur
job. Moreover,
diploma
must be more developed in

rising
sector

of the economy
.
like
services, marketing and ecology for example.
The diplomas
are older
compare to
the new job
and new sector of activity. The learning programs
must be refresh

This
different proposition
must be fight
the gap

between education and
labour market
and thus reduce the employment crisis.
</body>
</document>

Figure 53 : Exemple d'un fichier annoté (3/3)

Table des matières

Index :

Cartographie des erreurs en anglais L2 : vers une typologie intégrant système et texte

Résumé
L'objectif principal de ce travail est d'explorer la frontière entre les erreurs grammaticales d'une part et les erreurs textuelles d'autre part, dans les productions écrites des étudiants francophones rédigeant en anglais langue étrangère (L2) à l'université. Pour ce faire, un corpus de textes d'apprenants en anglais L2 a été recueilli et annoté par le biais de plusieurs schémas d'annotation. Le premier schéma d'annotation est issu de l'UAM CorpusTool, un logiciel qui fournit une taxonomie d'erreurs intégrée. Les premières annotations ont été croisées avec d'autres annotations issues des métafonctions sémantiques que nous avons établies, en nous appuyant sur la linguistique systémique fonctionnelle.
En plus de fournir des statistiques en termes de fréquence d'occurrence des erreurs spécifiques chez les apprenants francophones, le croisement des schémas a permis d'identifier certaines valeurs proprement phraséologique, sémantique et textuelle qui semblent poser des problèmes particulièrement épineux. A ce titre, une classification de ce que nous avons appelé des erreurs d'acceptabilité textuelle a été établie, dans le but notamment d'avoir une vue globale sur les erreurs identifiables à ce niveau d'analyse. En bref, le présent travail retrace donc le cheminement de l'ensemble de notre thèse de ses débuts conceptuels jusqu'à la proposition d'un modèle explicatif permettant d'établir la description de toute occurrence erronée identifiée en langue étrangère – qu'elle soit notamment grammaticale (c'est-à-dire, imputable au système linguistique) ou textuelle (c'est-à-dire, imputable au texte).
Mots clés : corpus d'apprenants, erreurs d'apprenant, linguistique systémique fonctionnelle, grammaticalité, acceptabilité, anglais de spécialité

Mapping English L2 errors: an integrated system and textual approach

Abstract
The main objective of this study is to try and pinpoint the frontier between grammatical (or sentence-level) errors on the one hand and textual errors on the other in university student essays. Accordingly, a corpus of English L2 learner texts, written by French learners, was collected and annotated using several annotation schemes. The first annotation scheme used is based on a model from the UAM CorpusTool software package, which provided us with an integrated error taxonomy. The annotations obtained were then cross-analyzed using the semantic metafunctions identified in systemic functional linguistics.
In addition to providing statistics in terms of specific error frequency, our cross analysis has identified some areas that appear to pose particularly difficult problems, i.e. phraseology, and certain semantic and textual constructions. A classification of what we have called textual acceptability errors has thus been established. In short, the thesis begins with an examination of conceptual issues and ends with the proposal for an explanatory model that can describe erroneous occurrences identified in a foreign language – whether they are grammatical (i.e., linked to the language system itself) or textual (i.e. linked to the text) in nature.
Keywords: learner corpus, learner errors, grammaticality, acceptability, English for Specific Purposes, systemic functional linguistics

UNIVERSITE SORBONNE NOUVELLE - PARIS 3
École doctorale 268 « langage et langues »
UMR 8094 - Langues, Textes, Traitements informatiques, Cognition (LATTICE) CNRS/ENS
Centre Bièvre, 1– 5 rue Censier, 75005 Paris